TRABAJO DE FIN DE GRADO

Grado en Estadística

Título: Estudio de asociación entre polimorfismos genéticos y fuerza muscular de estudiantes

Autor: Blanca Rius Sansalvador

Director: Jan Graffelman

Departamento: Estadística i Investigació Operativa (UPC)

Convocatoria: Juny 2019

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Mi más sincero agradecimiento a las personas que han hecho posible la realización y la culminación de este trabajo. En primer lugar, y en especial, a mi tutor Jan Graffelman por su tiempo, consejos y la orientación recibida. También quiero dar las gracias mis padres Silvia y Juan Manuel, mi hermana Manne y mis amigos por tanta paciencia. Para acabar, no quiero dejar de lado el apoyo recibido por parte de mi pareja Francesc. Gracias a todos.

Resumen y palabras clave

La fuerza muscular es primordial para la salud y actividad humana. Sin embargo, el conocimiento que se tiene de los factores genéticos que influyen en el tamaño y fuerza muscular es escaso. El proyecto *Functional single nucleotide polymorphisms Associated with Muscle Size and Strength* (FAMuSS) es un estudio cuyo objetivo es averiguar la asociación de 225 polimorfismos genéticos con el tamaño y la fuerza muscular antes y después de 12 semanas de entrenamiento de fuerza de 1396 estudiantes de distintas etnias. Se practica una exhaustiva depuración de los datos tratando los *missings* y valorando los polimorfismos mediante pruebas como el equilibrio de Hardy-Weinberg, el estudio de la frecuencia del alelo menos común, análisis descriptivos y el análisis de componentes principales. Se investigan los factores genéticos potencialmente relacionados con la ganancia de fuerza muscular mediante modelos lineales uni y multivariantes usando polimorfismos y haplotipos en esquemas genéticos aditivos, recesivos, dominantes y codominantes. Los métodos empleados detectan algunos polifmorfismos y haplotipos asociados con la ganancia de la fuerza muscular.

Palabras clave: análisis de componentes principales, datos faltantes, equilibrio de Hardy-Weinberg, estudio FAMuSS, frecuencia alélica, fuerza muscular, haplotipos, marcador genético, modelización, polimorfismo de nucleótido simple (SNP).

Clasificación AMS: 92D10 *Genetics*, 62J05 *Linear regression* y 62P10 *Applications to biology and medical sciences*.

Abstract

Muscular strength is critically important to health and human activity. The knowledge we have of the genetic factors that influence the size, strength and response to muscular exercise is small. The project "Functional single nucleotide polymorphisms Associated with Muscle Size and Strength" (FAMuSS) is a study funded by National Institutes of Health (NIH) in the United States and carried out in different centers to investigate the influence of the 225 selected genetic polymorphisms on the size and muscular strength before and after 12 weeks of strength training. The dominant arm has not been trained during the study, so it will be useful as control. The muscle size measurement, the isometric and dynamic force will be looked at through standardized tests and magnetic resonances. There are 346 variables, of which 225 of them are genetic polymorphisms, 76 are of muscular performance, 14 are physiological and 3 demographic variables. Data was collected from 1396 individuals, of which 632 were studied after the data was cleaned up. The subjects are students, all between 18 and 40 years of different ethnic groups who trained their non-dominant arm for 12 weeks. The database was exhaustively curated by analysing missing data, testing genetic polymorphisms for Hardy-Weinberg equilibrium, studying the distribution of the minor allele frequency, some descriptive analysis and principal component analysis. Genetic factors potentially related to a gain of muscle strength were studied by using uni and multivariate linear models using both polymorphisms and haplotypes in additive, recessive, dominant or co-dominant form. The applied methodology detected some polymorphisms and haplotypes associated with muscle strength gain.

Keywords: allele frequency, FAMuSS study, genetic marker, haplotypes, Hardy-Weinberg equilibrium, missing data, modeling, muscular strength, Principal Component Analysis, single nucleotide polymorphism.

Tabla de contenido

I. Introducción

El tejido muscular cubre aproximadamente el 30% de la masa corporal y responde rápida y eficientemente a los estímulos ambientales y al ejercicio físico. Tiene una influencia importante en las capacidades funcionales y ha sido asociada especialmente de manera positiva con el rendimiento deportivo (ACSM, 2009). La fuerza física también es primordial para la capacidad funcional de otros tejidos del organismo, como el mantenimiento de la densidad ósea. Se sabe que los estímulos del entorno son factores determinantes de la fuerza y tamaño muscular y, a pesar de la importancia para la actividad y salud humanas, se conoce poco sobre los factores genéticos (SNPs, polimorfismos de nucleótido único) que influyen en el tamaño muscular, la fuerza y la respuesta al ejercicio (Stewart y Rittweger, 2006). Se estima que los factores hereditarios pueden corresponder del 44 al 58% de las variaciones inter-individuo en la fuerza y tamaño muscular (Beunen y Thomis, 2004). El estudio FAMuSS *(Functional SNPs Associated with Muscle Size and Strength)* fue diseñado para identificar factores genéticos. Es un estudio realizado en múltiples centros, con financiación del *National Institutes of Health en Estados Unidos* (NIH), para buscar la influencia de polimorfismos genéticos en la medida muscular y fuerza antes y después de un entrenamiento de resistencia muscular.

Los factores genéticos que influyen en el tamaño muscular y en el carácter de los animales de granja están bien estudiados a causa de la importancia económica de la carne. El gen de la *miostatina* ha sido identificado como el gen principal que afecta al tamaño muscular y calidad del ganado. Adicionalmente, se han identificado dos genes más, el gen receptor de *rianodina* y el gen *IGF-2*, que afectan a la medida y calidad muscular en cerdos. Actualmente no hay evidencia de efectos genéticos similares en humanos (Thompson, 2004; Foulkes, 2009).

En el estudio FAMuSS, más de mil hombres y mujeres de 18 a 40 años han entrenado de la misma manera su brazo no dominante durante 12 semanas. Se ha tomado la medida muscular y la fuerza isométrica y dinámica antes y después del entrenamiento. Se investigará el efecto de unos genes candidatos implicados en el desarrollo de la fuerza y la fisiología muscular. Por último, se tratará de ver si existe relación estadística entre polimorfismos y el tamaño y fuerza musculares antes y después del ejercicio.

Este análisis de los datos del estudio FAMuSS ayudará a identificar los factores genéticos asociados a la actividad muscular y respuesta al ejercicio. Debería ayudar a ser capaces de predecir la respuesta de los individuos al ejercicio, a entender también la fisiología muscular y a identificar los sujetos susceptibles a perder músculo bajo ciertos desafíos ambientales. Este estudio podría ayudar a desarrollar agentes farmacológicos capaces de mantener el tamaño y la actividad muscular, una gran aplicación para el ámbito deportivo. Esta base de datos no ha sido todavía analizada exhaustivamente como se pretende hacer en este trabajo, ya que los pocos análisis anteriores existentes no han llegado a conclusiones prácticas.

La base de datos del estudio FAMuSS está descrita en varios artículos como el de Paul D. Thomson (*Functional Polymorphisms Associated with Human Muscle Size and Strength*, 2004) y en el libro de Andrea S. Foulkes (*Applied Statistical Genetics with R*, 2009) y es de carácter público (http://www.biostat.umn.edu/~cavanr/FMS_data.txt).

En estre trabajo, se especifica, en primer lugar, cuál ha sido la metodología utilizada para la realización del estudio: los pasos que se han seguido, el procedimiento en la depuración de los datos, las pruebas que se han utilizado y con qué nivel de significación entre otras. En segundo lugar, se ha realizado el diseño del estudio, donde se explican qué individuos han sido excluidos del estudio, la metodología de los test realizados, el programa de entrenamiento físico, los métodos de descubrimiento y genotipado de SNP, las limitaciones del estudio, entre otros aspectos. A continuación, se describe la base de datos, donde se distinguen las variables demográficas, genéticas, de rendimiento muscular y fisiológicas. Seguidamente, se procede a la depuración de la base de datos, donde se realiza además una selección de variables e individuos de interés. En el siguiente apartado del trabajo se incluyen gráficos sencillos y alguna prueba de grupo de variables seleccionadas como más importantes. El siguiente capítulo de la memoria está dedicado a la exploración de los datos genéticos: se recodifican los niveles de los polimorfismos, se imputan los datos faltantes, se eliminan los SNPs monomórficos y se realiza un análisis de componentes principales. Se han seleccionado las dos variables de rendimiento muscular más representativas y se ha procedido a su modelización con las covariables más importantes. Se han considerado 4 modelos distintos para cada variable respuesta. Se ha realizado un análisis más exhausivo del gen "resistin" mediante la estimación de haplotipos para las personas de raza caucásica donde se han

realizado modelos univariantes y multivariantes. Por último, se incluye un apartado de conclusiones, el glosario donde consultar vocabulario, la bibliografía usada y listas de siglas, figuras y tablas. En el anexo se adjunta el código de R utilizado en la realización del estudio.

II. Metodología

Uno de los problemas más importantes que se ha tenido a lo largo del estudio es la gran cantidad de datos faltantes que tenía la base de datos original (casi 40%). Se han eliminado los individuos a los cuales se les había realizado el estudio en los últimos 5 trimestres o que tenían esta variable como dato faltante y cuyo porcentaje de datos faltantes era superior al 60%. Se han seleccionado para la base de datos final los individuos y variables que tenían únicamente menos de un 50% de *missings*. En este apartado se han utilizado únicamente métodos gráficos.

Se han realizado estadísticas descriptivas sencillas, así como gráficos de barras, *boxplots* y gráficos de pastel para la descripción de variables. Se han ejecutado también contrastes de hipótesis como la prueba de t de Student y la prueba F de Fisher. Se ha calculado y graficado también la frecuencia del alelo menos común por razas y para el total de los datos. Para ver si se cumplía el equilibrio de Hardy-Weinberg se ha recurrido a contrastes de hipótesis para las frecuencias alélicas. El nivel de significación utilizado en todos los test es 0.05. Cuando se han realizado diversos test se ha usado la Corrección de Bonferroni para evitar el falso positivo así como el *False Discovery Rate* para asegurar que hay un máximo de un 5% de error de tipo I.

Para preparar la exploración de los datos genéticos, éstos se han recodificado con un código numérico (*0* para el homocigoto más frecuente, *1* para los heterocigotos, *2* para el homocigoto menos frecuente y *NA* para los datos faltantes) para facilitar el estudio. Se ha considerado que los datos faltantes eran completamente aleatorios (MCAR) para la imputación de *missings*. Se han imputado calculando la proporción en la que aparece cada nivel del factor para cada polimorfismo y se ha obtenido una muestra aleatoria del número de datos faltantes con las proporciones calculadas. Se han eliminado los 11 SNPs monomórficos que se tenían en la base de datos original. Se ha realizado un análisis de componentes principales para analizar las variables genéticas de manera visual y ver si se tenían distintas poblaciones.

Mediante modelos lineales se ha tratado de averiguar qué covariables son más influyentes en la ganancia del test de fuerza y del test de máxima repetición. El método de selección de variables usado es el Criterio de Información de Akaike. Una vez seleccionadas las covariables más importantes se ha contrastado si cada parámetro asociado era o no distinto de 0. Seguidamente, se ha añadido al modelo con las covariables cada SNP y se ha visto cuales son los más significativos. Se han considerado 4 modelos para cada variable respuesta, ya que se ha tenido en cuenta el esquema que pueden seguir los polimorfismos: aditivo, recesivo, dominante o codominante. La validación de los modelos seleccionados se ha realizado gráficamente. Se ha utilizado el método ANOVA para comparar modelos cuando unos niveles de las variables salían significativos y otros no. Se ha realizado una comparación de los SNP significativos en los modelos y para cada variable respuesta.

Se ha dedicado un apartado a la examinación más exhaustiva del gen *resistin* para los individuos de raza caucásica mediante la generación de haplotipos para disminuir el falso positivo por la gran cantidad de variables testeadas que se tienen. Se ha escogido este gen porque hay 4 SNPs asociados al mismo que han salido significativos. Se han preparado los datos, se han estimado las frecuencias alélicas y se ha asignado la más probable a cada individuo. Se ha considerado el modelo univariante para la ganancia de fuerza fuerza isométrica y se ha analizado exhaustivamente el haplotipo que parecía ser el más influyente. La ganancia de fuerza en el test de repetición máxima también ha sido analizada mediante modelos lineales. En los dos casos, primero se ha realizado un modelo con los diplotipos más frecuentes y después otro modelo añadiendo también los diplotipos minoritarios, se quería observar cuál era su comportamiento. Por último, se ha realizado un modelo multivariante con las dos variables respuesta y el grupo de covariables seleccionado mediante el p-valor resultante de la función MANOVA. El valor crítico que se ha considerado en todo el estudio es 0.05.

Todos los test y gráficos han sido realizados en lenguaje R y la memoria del estudio en R Markdown.

III. Diseño del estudio

Se ha medido la fuerza dinámica e isométrica de la musculatura de flexores de los antebrazos antes y después de 12 semanas de entrenamiento, así como la sección transversal de la parte superior del brazo de más de 1000 sujetos entre 18 y 40 años. Los sujetos de más de 40 años se han excluido con tal de evitar estudiar hombres que han sufrido una reducción en los niveles de testosterona.

Se realizan una serie de test para medir la cantidad de fuerza. La sección transversal máxima de bíceps y tríceps, en cambio, ha sido medida mediante resonancias magnéticas. Las muestras de ADN han sido obtenidas de los glóbulos blancos de muestras de sangre tomadas de los sujetos.

Los test antes del entrenamiento se realizan en 3 días mientras que los test después del entrenamiento en 2.

Hombres y mujeres son excluidos si:

- tienen < 18 o > 40 años
- usan medicaciones que se sabe que afectan a la medida muscular como corticosteroides
- tienen alguna restricción en actividades físicas
- tienen alguna condición médica crónica, como diabetes
- llevan implantes metálicos en brazos, ojos, cabeza, cuello o corazón, ya que tienen prohibidas las resonancias magnéticas
- han practicado entrenamientos de fuerza o trabajos que requieran el uso de los brazos repetitivamente en los 12 meses anteriores
- consumen, de media, más de dos bebidas alcohólicas diariamente
- toman suplementos alimenticios para aumentar el tamaño o fuerza muscular o para ganar peso, como proteínas, creatina o precursores androgénicos

Test isómetrico de fuerza de bíceps

La fuerza isométrica de bíceps de cada brazo ha sido testeada antes de las 12 semanas de entrenamiento de fuerza utilizando un banco especialmente construido para el estudio. Según la literatura, los valores del coeficiente de fiabilidad intraclase (R) para la flexión del codo a 90^o están entre 0.95 y 0.99. Los nueve centros implicados en el proyecto tienen los mismos bancos y llevan a cabo el test de la misma manera. Como se ha dicho, las pruebas antes del entrenamiento se hacen en 3 días, separados por menos de 2. En cambio, los test de después del entrenamiento se llevan a cabo en 2 días, el primero justo antes del último entrenamiento y el segundo 48h después del último entrenamiento.

En cada uno de los días de test, se realizan tres contracciones isométricas máximas con cada brazo. A fin de obtener tres valores de fuerza consistentes, se realizan hasta dos contracciones más para ver si valor de la última contracción se desvía más de 5 libras de los otros dos valores. La media de los resultados obtenidos en el segundo y el tercer día de test antes del entrenamiento se utilizará como medida basal y los resultados obtenidos 48 horas después serán escogidos como criterio de mesura.

Para el test, el brazo será posicionado en el banco con el codo fijo en un ángulo de 90^o. Dicho brazo se alineará con el antebrazo del sujeto, la muñeca del sujeto se colocará en un soporte acolchado, el codo del sujeto se alineará con el centro de rotación del brazo. El sujeto deberá empujar el antebrazo contra una resistencia fija situada en un ángulo de 45^o con su fuerza máxima (Figura 1). Para evitar que se trabaje con otros músculos, se posicionará al sujeto con su pecho contra el banco y las piernas rectas y con solamente los talones en el suelo. El otro brazo que no se esté testeando irá apoyado en las piernas. Se valorarán tres contracciones máximas. Cada contracción durará 3 segundos y se permitirá un descanso de 1 minuto entre contracciones. El promedio de la fuerza máxima producida durante las tres contracciones es la medida que se utilizará.

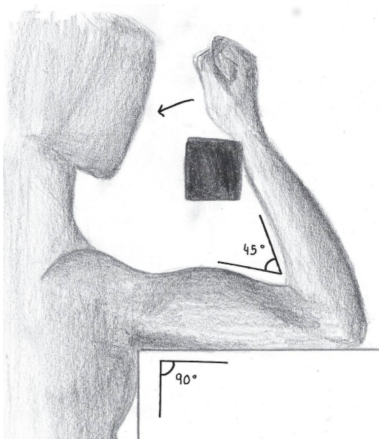

Figura 1: Recreación test isométrico de fuerza de bíceps.

Una repetición de fuerza máxima de bíceps (1RM)

La fuerza dinámica de los músculos flexores del codo de cada brazo se evaluará determinando la máxima cantidad de peso con el que un sujeto puede realizar una repetición del ejercicio de fuerza de bíceps explicado anteriormente. Este test se realizará antes y después del entrenamiento de fuerza. A diferencia de las pruebas de resistencia isométrica, el nivel basal de repetición máxima se completará en 1 día. Para evitar que la fatiga muscular afecte a los resultados, esta prueba se realizará después de las pruebas isométricas.

Evaluación de la ganancia de fuerza del bíceps

Debido a que las puntuaciones de fuerza no son perfectamente proporcionales a las diferencias intersectoriales en el tamaño corporal, las puntuaciones de fuerza se escalarán alométricamente al dividir el valor registrado entre la masa corporal. La ganancia de fuerza se determinará como la diferencia entre la medición previa y la medida normalizada. Este procedimiento se aplicará por separado a las puntuaciones de fuerza isométrica y a la repetición máxima de fuerza de bíceps. Se espera que haya individuos con una gran ganancia de fuerza mientras otros con una ganancia más pequeña pero con un tamaño corporal similar. Debido a las diferencias en la estatura corporal y hormonal del cada género, las personas con una ganancia de fuerza grande y pequeña se determinarán por separado para mujeres y hombres.

Medición del área transversal muscular

Con tal de evitar el falso aumento en las mediciones de la imagen por resonancia magnética (IRM) a causa de la inflamación del músculo, esta se realizará antes o entre 24 y 48 horas después del test. Esto garantiza que se eviten efectos temporales como los cambios en la cantidad de agua en el cuerpo que pueden hacer que varíe el tamaño muscular.

Los datos de medición del área transversal muscular posteriores al entrenamiento se compararán con los valores previos para determinar los aumentos inducidos por el mismo.

Las imágenes de resonancia magnética previas y posteriores al entrenamiento se obtendrán por separado en los brazos dominantes (no entrenados) y no dominantes (entrenados), lo que permitirá que el brazo dominante actúe como control.

Debido a que las imágenes de RM se recopilarán antes y después del entrenamiento, es importante que la posición de cada sujeto dentro del imán de RM se reproduzca de la misma manera en las dos medidas para evitar errores de registro. Para lograr esto, la circunferencia máxima de la parte superior del brazo se medirá con una cinta aislante. El brazo se colocará a 90º respecto el cuerpo y se flexionará 90º en el codo. El bíceps se contraerá al máximo para esta medida (Figura 2).

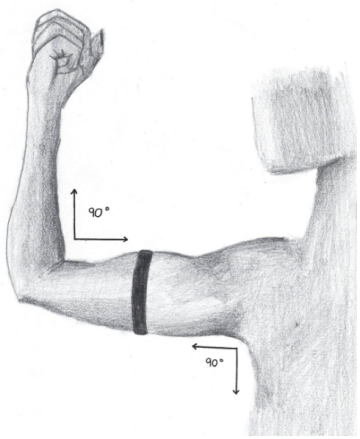

Figura 2: Recreación de la medición de la circunferencia máxima.

La ubicación de la circunferencia máxima (POM), o punto de medida, se marcará en la piel del sujeto antes de la medición.

Los sujetos tendrán ambos brazos escaneados en posición supina, con el brazo de interés a su lado y el centro del brazo lo más cerca posible del isocentro magnético del escáner, con las palmas hacia arriba. La mano estará boca arriba y pegada con cinta adhesiva en la superficie del lecho del escáner, y la POM se centrará en la luz de alineación de la IRM.

Según la literatura, se obtendrá una imagen de explorador coronal (6-9 cortes) para ubicar el eje largo del húmero, seguido de una imagen de exploración sagital (6 a 9 cortes) para alinear el octavo corte de la imagen axial / oblicua con el POM. A continuación, se obtendrán quince imágenes de gradiente degradadas en serie de cada brazo (TE = 1.9 s, TR = 200 ms, supresión de artefactos de flujo, ángulo de giro de 30º) utilizando el POM como el punto más central. Estos cortes de imagen axiales / oblicuos (es decir, perpendicular al húmero) comenzarán en la parte superior del brazo y avanzarán hacia el codo, de manera que el vientre del músculo se produzca en los cortes 8 y 9. Los cortes de imagen individuales tendrán un grosor de 16 mm con un espacio intercalado de 0 mm y una matriz de resolución de 256 × 192, campo de visión de 22 cm × 22 cm, NEX = 6. Este método muestra una longitud de 24 cm de cada brazo.

Las imágenes de MR de cada sitio de investigación se transferirán a la instalación central de imágenes de MR en la *Yale University* a través de un Disco Óptico Magneto o un *CD-ROM*. Las imágenes se analizarán utilizando un programa de visualización y procesamiento interactivo de diseño personalizado que funciona en Matlab (*The Math Works, Inc., Natick, MA*) que se ejecuta en un ordenador con sistema operativo *Windows*. Este software permite al usuario asignar regiones de interés (ROI) en un conjunto de imágenes mediante el rastreo de los bordes de la región con el ratón. El músculo es fácilmente identificable en las IRM y su CSA se medirá utilizando esta técnica de planimetría computarizada. Una vez que se define el ROI, el programa informa la cantidad de píxeles contenidos en el ROI seleccionado. Sobre la base de los datos de adquisición de IRM (es decir, el campo de visión y la resolución de la matriz), se calcula la CSA (cm^2) del ROI definido. Cuando el CSA anterior al entrenamiento se resta del CSA posterior al entrenamiento, el efecto del entrenamiento puede compararse entre los sujetos. Con el fin de optimizar la precisión del cálculo del tamaño muscular, se analizará un subconjunto de datos mediante análisis volumétrico. Al analizar los 15 cortes sucesivos a lo largo de la longitud escaneada del brazo superior, cada CSA puede multiplicarse por el grosor de corte conocido (1.6 cm) para producir un volumen de corte (cm^3). Los volúmenes de corte se sumarán a lo largo de la estructura anatómica de interés. Aunque el análisis volumétrico generalmente se considera más preciso que el análisis transversal, la cantidad de trabajo requerido para realizar el análisis volumétrico en 2000 estudios (1000 sujetos con dos medidas cada uno) es prohibitiva.

Programa de entrenamiento físico

Los sujetos se someterán gradualmente a un entrenamiento de fuerza supervisado progresivo de su brazo

no dominante en una de las instalaciones colaboradoras. La repetición de fuerza máxima medida durante las pruebas previas al entrenamiento se utilizará para estimar los pesos que se pueden levantar con 12, 8 y 6 repeticiones utilizando fórmulas estándar. Todos los ejercicios se realizarán únicamente con el brazo no dominante. Los ejercicios consistirán en el ejercicio de bíceps en el banco, ejercicio de concentración de bíceps, ejercicio de bíceps de pie, extensión de tríceps por encima de la cabeza y el retroceso del tríceps.

Todos los ejercicios se realizarán con un equipo estandarizado. Las sesiones de entrenamiento serán supervisadas y durarán aproximadamente entre 45 y 60 minutos. La progresión del ejercicio utilizará el siguiente protocolo de entrenamiento semanal:

- Semanas 1-4: 3 grupos de 12 repeticiones del peso asignado por las fórmulas para este número de repeticiones
- Semanas 5–9: 3 grupos de 8 repeticiones del peso asignado por las fórmulas para este número de repeticiones
- Semanas 10-12: 3 grupos de 6 repeticiones del peso asignado por las fórmulas para este número de repeticiones

Este protocolo se ha diseñado para aumentar el tamaño de bíceps y tríceps usando muchas repeticiones a baja intensidad al principio del entrenamiento y más fuerza con pocas repeticiones al final, aumentando la intensidad a medida que el entrenamiento progresa. Esta transición activa la masa muscular a través de un mayor reclutamiento de unidades motoras. El interés principal es entrenar los flexores del codo, pero también se entrenan los extensores del codo para equilibrar la fuerza muscular a través de la articulación.

Tiempo	Prueba previa	Semana de test							Inicio entrenamiento
Días		1	2	3	4	5	6	7	
Muestra ADN	← Cualquier día →								
Test isométrico		← 3 sesiones que empiezan cualquier día separadas por 24-48 horas →							
Test repetición máxima				← Después del tercer test isométrico →					
IRM	Cualquier día, antes del test isométrico o < 24 horas después y > 48 horas despues del test de máxima repetición								

Figura 3: Protocolo de las pruebas antes del entrenamiento

Tiempo	Último día de entrenamiento	Prueba semana			
Días		1	2	3	4
Test isomético	Primer test justo antes de l'última sesion de entrenamiento		Segundo test 48-72 horas antes de l'última sesión de entrenamiento		
Test repetición máxima			Después del segundo test isométrico		
IRM			>24 horas después del test isométrico y > 48 horas después del test de repetición máxima y < 96 horas antes de l'última sesión de entrenamiento		

Figura 4: Protocolo de las pruebas después del entrenamiento

Procedimientos de control dietético

Los sujetos mantendrán su ingesta dietética habitual a lo largo del estudio. No se reclutará a las personas que hayan complementado su dieta con proteínas adicionales o que hayan tomado algún suplemento dietético para desarrollar músculo o aumentar el peso (suplementos dietéticos que contengan proteínas, creatina o precursores androgénicos). No se analizarán los datos de los sujetos que han perdido cantidades significativas de peso durante el estudio.

Flebotomía

Las muestras de sangre (21 cc) para las determinaciones de ADN se obtendrán de los sujetos antes de comenzar la parte de entrenamiento con ejercicios del estudio.

Mediciones antropométricas

El peso corporal se registrará antes y después del entrenamiento utilizando una balanza médica.

La altura se determinará utilizando una cinta métrica montada en una pared.

Las circunferencias de la parte superior de los brazos se medirán dos días antes y dos días después de las 12 semanas de entrenamiento utilizando una cinta métrica no estirable. Todas las mediciones de la circunferencia del brazo se realizarán antes de la sesión de prueba de una repetición máxima para evitar la posibilidad de que la inflamación muscular de la prueba pueda afectar el tamaño del brazo.

Estandarización entre centros

Para controlar cualquier diferencia entre los centros, cada uno ha utilizado el mismo protocolo de entrenamiento. El material necesario para el ejercicio está comprado a los mismos fabricantes. Las técnicas de IRM, la fuerza, las mediciones antropométricas y los ejercicios del entrenamiento se grabaron en video al inicio del estudio. Se ha requerido que el personal de investigación de cada sitio revise los procedimientos grabados en video antes del inicio de cada grupo de entrenamiento. Todos los investigadores principales y el personal involucrado en las pruebas se han reunido dos veces al año, una para revisar el progreso del estudio y otra para revisar la estandarización de las técnicas de medición.

Métodos de descubrimiento y genotipado de SNP

La fase inicial de la sección de genotipado elegirá una serie de genes candidatos (aproximadamente 100 genes), en base a tres criterios:

1) datos preliminares de asociación con fuerza, tamaño o respuesta al ejercicio en publicaciones anteriores
2) proteínas que se sabe que están en las rutas bioquímicas involucradas en la respuesta muscular al ejercicio
3) genes que se ha demostrado que están fuertemente regulados transcripcionalmente durante el entrenamiento aeróbico o el ejercicio excéntrico.

Dada la lista priorizada de "genes candidatos", la segunda fase consistirá en identificar polimorfismos dentro de los genes candidatos para la genotipificación posterior en los participantes del estudio. Según la literatura, se han utilizado dos enfoques para identificar SNPs dentro de los candidatos; minería de bases de datos utilizando bases de datos existentes, y utilizando el propio "descubrimiento de SNPs" independiente utilizando un panel de 96 individuos étnicamente diversos (36 caucásicos, 29 afroamericanos, 26 hispanos, 5 asiáticos). La lógica del propio descubrimiento de SNP es que la gran mayoría de los SNPs existentes están en secuencia no codificada. Estos SNP no codificantes son útiles para estudios de vinculación genética pero no son útiles para las asociaciones funcionales planificadas en nuestro estudio. Además, se reconoce que la sensibilidad y especificidad de los recursos de SNPs existentes es bastante pobre, con un máximo de 40.7% de SNPs (errores de secuenciación y clonación) y solo 47% de sensibilidad para el SNP existente en las secuencias de codificación de genes. El descubrimiento de SNPs se realizará mediante la amplificación de todos los exones, los límites exón / intrón, las regiones no traducidas 5' y 3' y los promotores de genes seleccionados. Los productos de PCR se someterán a una prueba de detección de polimorfismos utilizando uno de los dos métodos siguientes: secuenciación directa automatizada de todos los productos de PCR en los 96 individuos con análisis de trazas Phred / Phrap o cromatografía líquida de alta presión desnaturalizante (sistema de onda transgenómica) y posterior secuenciación de heterodúplex. La genotipificación de los sujetos del estudio se realizará mediante secuenciación directa, sistema de genotipado de Nanogen, digestiones de fragmentos de restricción de productos de PCR o química de nucleasa 5 'fluorogénica (ensayo Taqman). Se ha considerado realizar el propio descubrimiento de SNPs en valores atípicos fenotípicos extraídos de los resultados de FAMuSS en lugar de hacerlo desde el panel de detección étnicamente diverso de 96 personas. Sin embargo, existía la necesidad pragmática de realizar el descubrimiento de SNPs basado en laboratorio en paralelo con el brazo de fenotipado del estudio, tanto por razones presupuestarias como para evitar retrasar la adquisición de datos de asociación. Además, es bastante común identificar "SNP privados", donde se observa un polimorfismo en solo uno o muy pocos individuos. El uso de valores atípicos fenotípicos de los resultados iniciales de FA-MuSS para el brazo de descubrimiento de SNP de nuestro estudio podría llevarnos a considerar un "SNP privado" como asociado significativamente con un rasgo fenotípico específico en un valor atípico fenotípico,

cuando en cambio, esto sería una asociación aleatoria. En el diseño del estudio, el descubrimiento de SNPs es independiente del fenotipado.

Limitaciones del estudio

Existen varias limitaciones para el diseño. La mayoría de los ensayos en múltiples centros reducen la variabilidad de la medición utilizando un laboratorio central para medir muestras biológicas e interpretar las imágenes obtenidas en los centros participantes. Las muestras genéticas y de suero en FAMuSS, así como los resultados de MRI, se determinarán en los laboratorios centrales. No obstante, los investigadores realizarán las mediciones de la fuerza muscular en los sitios de investigación, lo que aumentará la variabilidad de la medición. Para minimizar el efecto de la variabilidad de la medida en el objetivo principal del estudio, esta se reducirá al estandarizar las técnicas en un manual de estudio y al certificar a los investigadores en cada sitio.

El personal de investigación también se reunirá dos veces al año con una reunión dedicada a revisar y practicar las técnicas de medición.

La ingesta dietética no se controlará en el estudio porque se considera que se proporciona suficiente proteína en la dieta estadounidense habitual para respaldar las demandas metabólicas de crecimiento muscular que se esperan en el entrenamiento con un solo brazo. A los sujetos se les pedirá que mantengan su ingesta dietética habitual, y no se reclutarán sujetos con dietas inusuales. Para evitar el efecto de dietas especiales, no se reclutarán sujetos que usen suplementos nutricionales o ayudas ergogénicas. Se consideró la idea de tener constancia de la dieta diaria de cada individuo, pero no se incluyeron en el diseño final del estudio debido a la carga adicional del análisis y del sujeto. El peso corporal se utilizará para garantizar que los sujetos no restrinjan las calorías durante el estudio, y los valores atípicos que pierden cantidades significativas de peso durante el estudio no se incluirán en el análisis genético. Sin embargo, las manipulaciones calóricas o dietéticas más sutiles no informadas por los sujetos o detectadas por la pérdida de peso podrían afectar los resultados.

Los voluntarios en FAMuSS pueden no representar a la población general. Los sujetos deben estar dispuestos a participar en un entrenamiento con un solo brazo durante 12 semanas, un criterio que puede afectar a la población del estudio. También es posible que las personas con genes del músculo esquelético que mejoran el tamaño o la fuerza muscular no sientan la necesidad de participar en el entrenamiento físico. Esto no debería afectar a aquellos valores atípicos con bajo tamaño muscular y fuerza de base, pero podría restringir el reclutamiento de individuos con parámetros de referencia altos.

El objetivo final de FAMuSS es realizar un análisis genético de todos los genes musculares y examinar los SNPs identificados en todos los sujetos. Esto permitirá un examen de la distribución de SNPs en comparación con la distribución fenotípica de los músculos y proporcionará un tamaño de muestra suficiente para examinar la interacción de varios SNPs.

IV. Descripción de la base de datos

Los datos del estudio son públicos (http://www.biostat.umn.edu/~cavanr/FMS_data.txt) y acompañan el libro *Applied Statistical Genetics with R* (Foulkes, 2004).

La base de datos original tiene 1396 observaciones y 346 variables. Sin embargo, se verá ésta se reduce a causa de problemas que irán surgiendo a medida que transcurre el estudio.

Las variables se agrupan globalmente en cinco categorías: unas pocas demográficas, muchas variables genéticas (básicamente polimorfismos de nucleótido único, conocidos como SNPs), variables de rendimiento muscular, covariables fisiológicas y algunas variables relacionadas con el diseño del estudio.

Las principales variables del estudio se enumeran a continuación, tras haber descartado otras cuyo significado no quedaba claro en la documentación disponible. Principalmente eran variables de volumen muscular en las que no estaba claro a qué músculo referían y algunas variables fisiológicas.

4.1 Variables demográficas

- `Gender`: género. *Female* para femenino y *Male* para masculino.
- `Age`: edad en años.
- `Race`: raza. *Caucasian* para caucásica, *Hispanic* para hispánica, *African Am* para afroamericana, *Asian* para asiática, *Other* para otras.

4.2 Variables genéticas

Refieren a información de 225 SNPs de 76 genes distintos de los individuos. Codificados como *2* para el homocigoto menos frecuente, *0* para el más frecuente y *1* para los heterocigotos (por ejemplo, si se tienen 100 polimorfismos A/A, 150 A/T y 30 T/T, la recodificación sería A/A $\rightarrow$ 0, A/T $\rightarrow$ 1, T/T $\rightarrow$ 2). Los polimorfismos genotipados pertenecen a un conjunto de genes implicados en la fisiología muscular que forman un juego de genes candidatos a tener relación con las variables de rendimiento muscular.

Los 225 SNPs llevan un nombre indicando el gen al cual pertenecen. Algunos de los genes implicados en la fisiología muscular son: *atcn3*, *esr1*, *resistin*, *aktn1*... Todos los SNPs tienen un identificador compuesto por el gen al que pertenecen seguido de un identificador específico para el polimorfismo. Por ejemplo, *actn3_rs540874* se refiere a un polimorfismo en el gen *atcn3*, siendo *rs540874* el identificador propio del mismo. De la misma manera se expresan los polimorfismos: *actn3_r577x*, *actn3_rs1815739*, *actn3_1671064* del mismo gen o *esr1_rs1801132*, *esr1_rs1042717*, *esr1_rs2228480* del gen *esr1*. Por cuestiones de espacio no se enumeran los nombres de los 225 SNPs.

4.3 Variables de rendimiento muscular

Hay unas 76 variables de rendimiento muscular, algunas de ellas son promedio o diferencias de varias de mediciones repetidas. Hay algunas variables como `bi_D`, `bi_ND`, `tri_D` y `tri_ND` que se han creado durante el estudio con tal de facilitar algunos cálculos.

- **`V1_ND1, V1_ND2, V1_ND3:`** resultados de los tres test del día 1 antes del entrenamiento en el brazo no dominante.
- **`V1_ND_AVG:`** promedio de los tres resultados del día 1 antes del entrenamiento en el brazo no dominante.
- **`V2_ND1, V2_ND2, V2_ND3:`** resultados de los tres test del día 2 antes del entrenamiento en el brazo no dominante.
- **`V2_ND_AVG:`** promedio de los tres resultados del día 2 antes del entrenamiento en el brazo no dominante.

- **V3_ND1, V3_ND2, V3_ND3:** resultados de los tres test del día 3 antes del entrenamiento en el brazo no dominante.
- **V3_ND_AVG:** promedio de los tres resultados del día 3 antes del entrenamiento en el brazo no dominante.
- **V23_ND_AVG:** promedio de los dos resultados del día 2 y 3 antes del entrenamiento en el brazo no dominante.
- **V123_ND_AVG:** promedio de los tres resultados del día 1, 2 y 3 antes del entrenamiento en el brazo no dominante.
- **Post1_ND1, Post1_ND2, Post1_ND3:** resultados de los tres test del día 1 después del entrenamiento en el brazo no dominante.
- **Post1_ND_AVG:** promedio de los tres resultados del día 1 después del entrenamiento en el brazo no dominante.
- **Post2_ND1, Post2_ND2, Post2_ND3:** resultados de los tres test del día 2 después del entrenamiento en el brazo no dominante.
- **Post2_ND_AVG:** promedio de los tres resultados del día 2 después del entrenamiento en el brazo no dominante.
- **Post_ND_avg:** promedio de los resultados después del entrenamiento del brazo no dominante.
- **ND23_DIFF:** diferencia entre Post_ND_avg y V23_ND_AVG, promedio de las medidas de después del entrenamiento y de los días 2 y 3 antes del entrenamiento en el brazo no dominante. Podría decirse que es la fuerza que se ha ganado con el entrenamiento.
- **V1_D1, V1_D2, V1_D3:** resultados de los tres test del día 1 antes del entrenamiento en el brazo dominante.
- **V1_D_AVG:** promedio de los tres resultados del día 1 antes del entrenamiento en el brazo dominante.
- **V2_D1, V2_D2, V2_D3:** resultados de los tres test del día 2 antes del entrenamiento en el brazo dominante.
- **V2_D_AVG:** promedio de los tres resultados del día 2 antes del entrenamiento en el brazo dominante.
- **V3_D1, V3_D2, V3_D3:** resultados de los tres test del día 3 antes del entrenamiento en el brazo dominante.
- **V3_D_AVG:** promedio de los tres resultados del día 3 antes del entrenamiento en el brazo dominante.
- **V23_D_AVG:** promedio de los dos resultados del día 2 y 3 antes del entrenamiento en el brazo dominante.
- **V123_D_AVG:** promedio de los tres resultados del día 1, 2 y 3 antes del entrenamiento en el brazo dominante.
- **Post1_D1, Post1_D2, Post1_D3:** resultados de los tres test del día 1 después del entrenamiento en el brazo dominante.
- **Post1_D_AVG:** promedio de los tres resultados del día 1 después del entrenamiento en el brazo dominante.
- **Post2_D1, Post2_D2, Post2_D3:** resultados de los tres test del día 2 después del entrenamiento en el brazo dominante.
- **Post2_D_AVG**: promedio de los tres resultados del día 2 después del entrenamiento en el brazo dominante.
- **Post_D_avg:** promedio de los resultados después del entrenamiento del brazo dominante.
- **D23_DIFF:** diferencia entre Post_D_avg y V23_D_AVG, promedio de las medidas de después del entrenamiento y de los días 2 y 3 antes del entrenamiento en el brazo dominante. Podría decirse que es la fuerza que se ha ganado con el entrenamiento.
- **Pre_RBi_Avg:** media de la sección transversal del bíceps derecho antes del entrenamiento.
- **Pre_RTri_Avg:** media de la sección transversal del tríceps derecho antes del entrenamiento.
- **Pre_LBi_Avg:** media de la sección transversal del bíceps izquierdo antes del entrenamiento.

- **Pre_LTri_Avg:** media de la sección transversal del tríceps izquierdo antes del entrenamiento.
- **Post_RBi_Avg:** media de la sección transversal del bíceps derecho después del entrenamiento.
- **Post_RTri_Avg:** media de la sección transversal del tríceps derecho después del entrenamiento.
- **Post_LBi_Avg:** media de la sección transversal del bíceps izquierdo después del entrenamiento.
- **Post_LTri_Avg:** media de la sección transversal del tríceps izquierdo después del entrenamiento.
- **bi_ND**: diferencia media entre después y antes del ejercicio de la sección transversal del bíceps del brazo no dominante.
- **bi_D**: diferencia media entre después y antes del ejercicio de la sección transversal del bíceps del brazo dominante.
- **tri_ND**: diferencia media entre después y antes del ejercicio de la sección transversal del tríceps del brazo no dominante.
- **tri_D**: diferencia media entre después y antes del ejercicio de la sección transversal del tríceps del brazo dominante.
- **Pre_NDRM_Max:** repetición fuerza máxima de bíceps brazo no dominante antes del entrenamiento
- **Post_NDRM_Max:** repetición fuerza máxima de bíceps brazo no dominante después del entrenamiento
- **NDRM_DIFF:** diferencia repetición máxima de bíceps después y antes del entrenamiento del brazo no dominante
- **Pre_DRM_Max:** repetición fuerza máxima de bíceps brazo dominante antes del entrenamiento
- **Post_DRM_Max:** repetición fuerza máxima de bíceps brazo no dominante después del entrenamiento
- **DRM_DIFF:** diferencia repetición máxima de bíceps después y antes del entrenamiento del brazo dominante

4.4 Variables fisiológicas

- **Pre.weight**: peso antes de empezar el estudio en kilogramos.
- **Pre.height**: altura antes de empezar el estudio en centímetros.
- **pre.BMI**: índice de masa corporal antes de empezar el estudio.
- **SBP**: presión sanguínea sistólica.
- **DBP**: presión sanguínea diastólica.
- **Dom.Arm**: mano dominante. *Left* para zurdos y *Right* para diestros.
- **Post.weight**: peso antes al acabar el estudio en kilogramos.
- **Post.Height**: altura al acabar el estudio en centímetros.
- **Calc.post.BMI**: índice de masa corporal al acabar el estudio.
- **HDL_C:** lipoproteínas de alta densidad, colesterol "bueno"
- **VLDL_TG:** lipoproteínas de muy baja densidad.
- **LDL_C:** lipoproteínas de baja densidad.
- **CRP:** nivel de proteína C reactiva.
- **Mean_BP**: media presión sanguínea.
- **HOMA:** índice que permite precisar un valor numérico expresivo de resistencia insulínica.

4.5 Otras variables

- `Center`: Centro donde se han llevado a cabo las pruebas. "FA" para *Florida Atlantic University*, "FL" para *University of Central Florida*, "HH" para Hartfort Hospital, "IR" para *Dublin City University* (Irlanda), "MA" para *University of Massachusetts*, "MI" para *Central Michigan University*, "UC" para *University of Connecticut* y "WV" para *University of West Virginia*.
- `Term`: Periodo de tiempo en el que ha sido estudiado el individuo. "02-1" para el primer trimestre de 2002, "02-2" para el segundo trimestre de 2002, "02-3" para el tercer trimestre de 2002, "03-1" para el primer trimestre de 2003, "03-2" para el segundo trimestre de 2003, "03-3" para el tercer trimestre de 2003, "04-1" para el primer trimestre de 2004, "04-2" para el segundo trimestre de 2004, "04-3" para el tercer trimestre de 2004 y "05-1" para el primer trimestre de 2005.

V. Selección de variables e individuos

A la vista de la gran cantidad de información disponible, se ha hecho una selección previa de variables de interés para el análisis descrito en esta memoria, escogiendo las variables `ND23_DIFF` (ganancia de fuerza isométrica de bíceps) y `NDRM_DIFF` (ganancia en el test de repetición máxima) como variables respuesta de rendimiento muscular, todos los polimorfismos como variables explicativas potenciales y covariables relevancia conocida como sexo, edad y raza.

Se ha recategorizado la variable `CRP` (nivel de proteína C reactiva), que tenía 24 categorías, la mayoría de 1 o 2 individuos, en 3 categorías. Había un solo individuo de raza aborigen y se ha incluido en la categoría *other*. Se han cambiado las unidades de algunas variables con tal de facilitar su interpretación. Se ha pasado de libras a kilogramos y de pulgadas a centímetros. Se han descartado los individuos que han perdido más de 10kg en las 12 semanas.

Seguidamente, se procede al estudio de los datos faltantes.

5.1 Porcentaje de datos faltantes por variable

Se calcula el porcentaje *missings* por variable y se representa en un histograma su distribución (Figura 5).

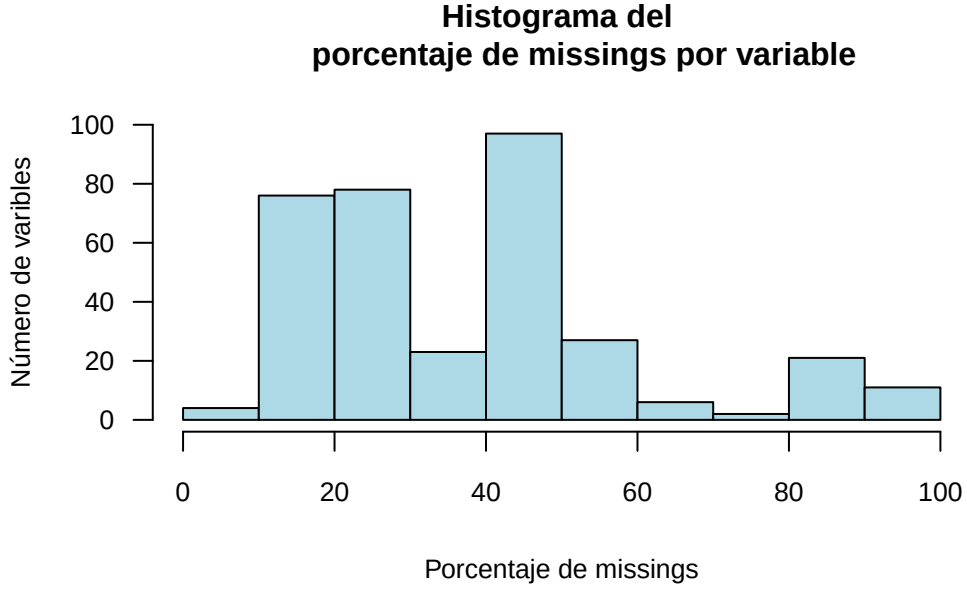

Figura 5: Histograma del porcentaje de missings de la base de datos original por variable.

Se puede observar que hay un grupo de variables con más de un 80% de *missings*, un gran grupo de variables que ronda el 20% y otro entre el 40% y 50%. Se ha observado que hay 32 variables con más de un 80% de *missings* y, se ha visto que la mayoría son SNPs.

5.2 Porcentaje de datos faltantes por individuo

Se calcula el porcentaje *missings* por individuo y se representa en un histograma su distribución (Figura 6)

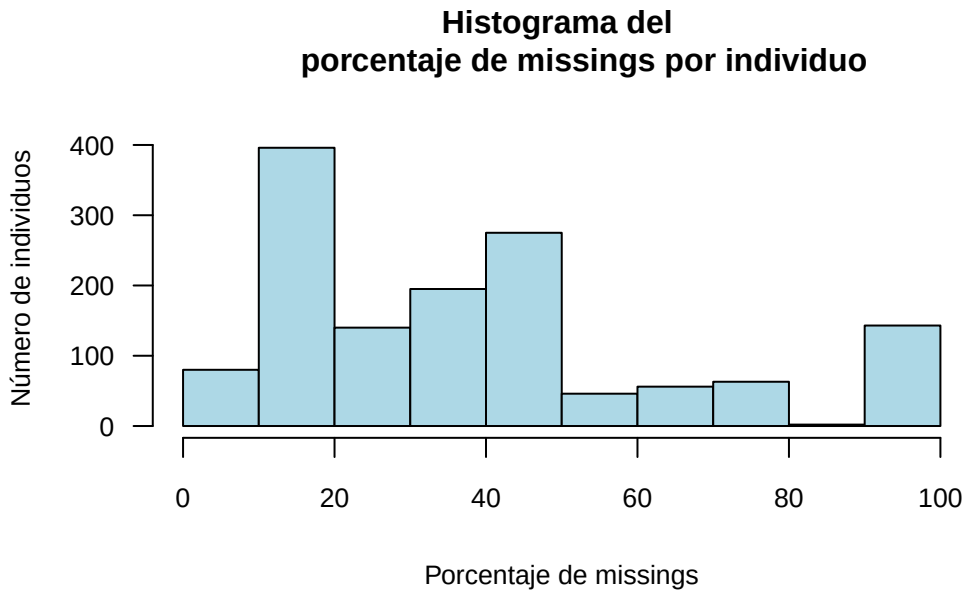

Figura 6: Histograma del porcentaje de missings de la base de datos original por individuo.

Se puede observar que hay un grupo de individuos con más de un 90% de *missings*, un gran grupo de variables que ronda el 10-20%. El resto parece estar entre el 20-50%. Se encuentran 145 individuos tienen más de un 80% de datos faltantes.

5.3 Porcentaje de datos faltantes total

Hay un 38.62% de *missings* en la base de datos original. Es un porcentaje bastante alto, en los siguientes apartados se tratará de solucionar el problema.

5.4 Eliminación de variables e individuos con muchos *missings*

Datos faltantes por trimestre

Se sospecha que el resto de los datos faltantes que se tienen no son aleatorios y la que variable `Term` está muy relacionada con la gran cantidad de ellos que se tiene. Se calculará el número de datos faltantes por cada uno de los trimestres. Se representarán gráficamente dichos porcentajes estratificando por trimestre para ver si existe algún tipo de patrón (Figura 7).

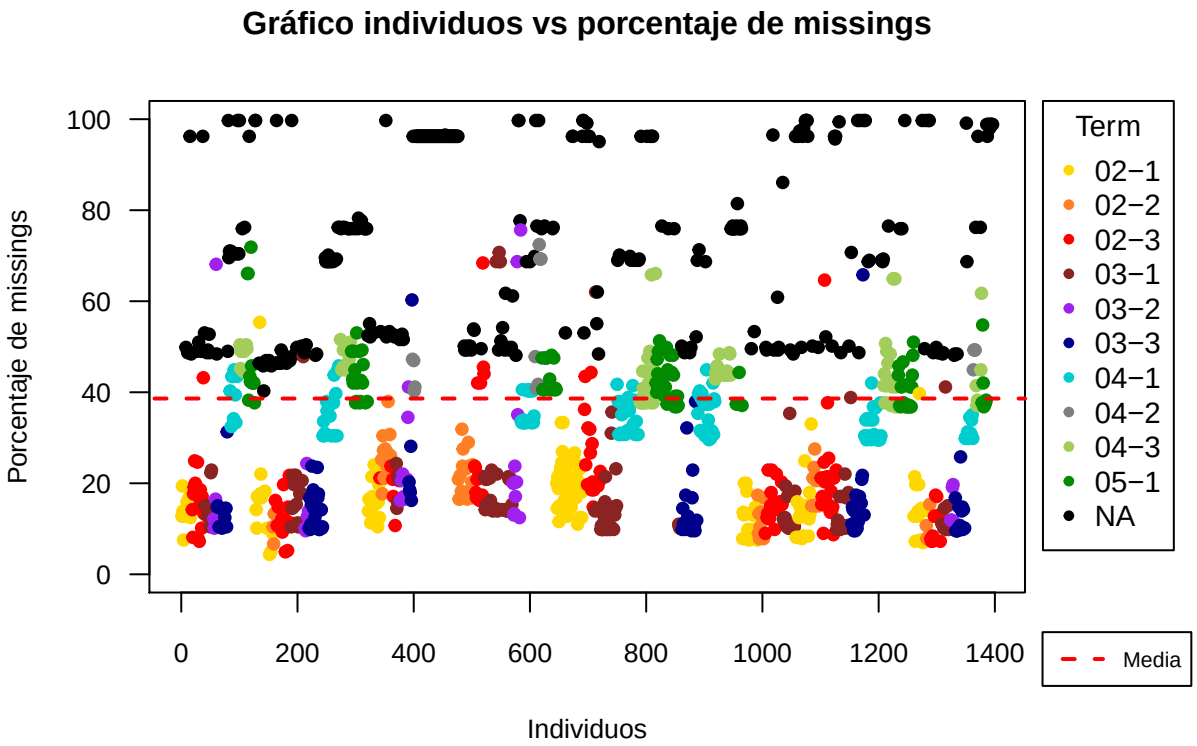

Figura 7: Individuos vs. porcentaje de missings.

Se puede observar que parece que los datos faltantes estén agrupados y que haya trimestres en los cuales la mayoría de los puntos se encuentran por encima de la recta que denota el promedio de *missings* y otros por debajo.

Con el diagrama de barras de la Figura 8 se ve de manera más clara que sí es cierto que algunos trimestres tienen una cantidad de *missings* significativamente mayor al resto.

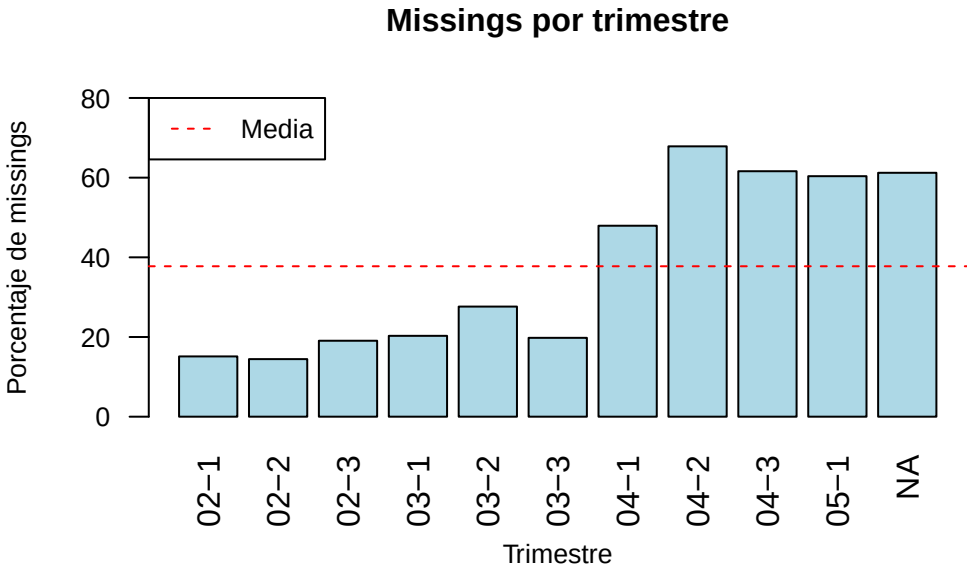

Figura 8: Porcentaje de datos faltantes por timestre.

En el gráfico de barras anterior puede observarse que a partir del trimestre "03-3" el número de datos faltantes es mucho más elevado, pasa del 15-25% a 50-70%. Para los individuos que tienen `Term` como dato faltante, el porcentaje de *missings* es de alrededor del 60%.

Por lo tanto, se decide que se estudiarán solamente los 6 primeros trimestres.

Porcentaje de datos faltantes por variable

Se había visto que la mayoría de *missings* no eran esporádicos, sino que están concentrados en algunas variables y en algunos individuos. Se estudia cómo es su distribución después de la extracción de algunos trimestres de la base de datos.

Se vuelve a calcular el porcentaje *missings* por variable y se representa en un histograma su distribución (Figura 9).

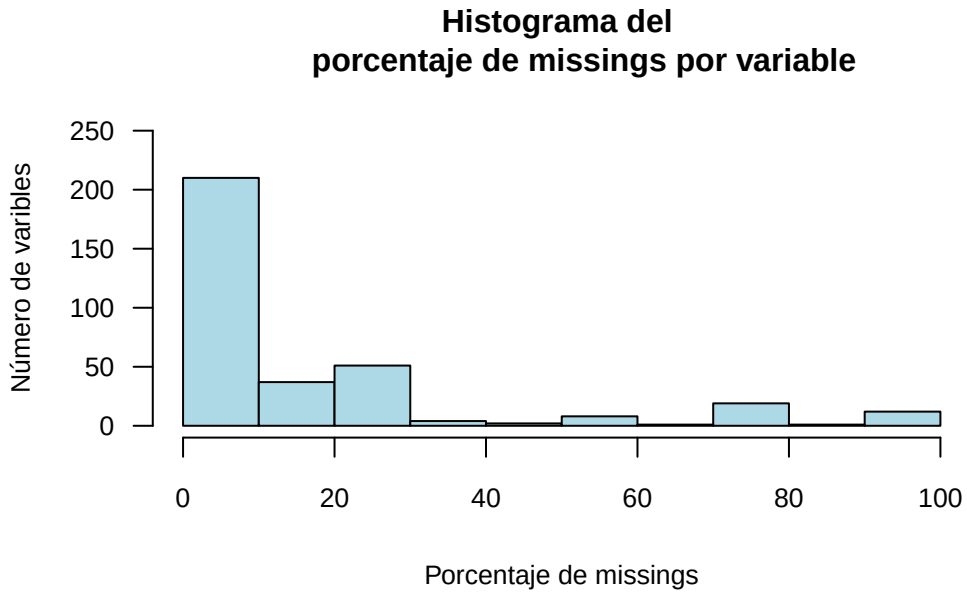

Figura 9: Histograma del porcentaje de datos faltantes por variable.

Hay 41 variables que tienen un porcentaje de datos faltantes que excede el 50%, porlo tanto, se eliminan. Esto es debido a que se considera que no aportan casi información.

Porcentaje de datos faltantes por individuo

Se vuelve a calcular el porcentaje *missings* por individuo y se representa en un histograma su distribución (Figura 10)

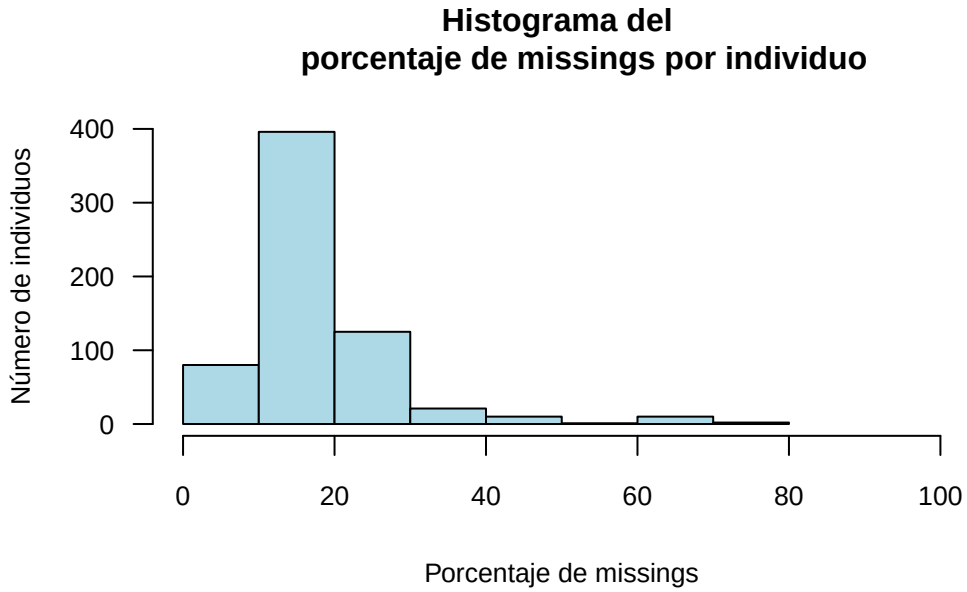

Figura 10: Histograma del porcentaje de datos faltantes por individuo.

Se observa que el porcentaje de *missings* por individuo conseguido con la eliminación de algunos trimestres es muy inferior al anterior, no obstante, también se procederá a eliminar los individuos con más de un 50% de datos faltantes siguiendo el mismo criterio que en el apartado anterior.

Se han eliminado un total de 764 individuos y 41 variables. La mayoría de las variables eliminadas son datos genéticos. Restan 632 individuos y 304 variables para el estudio, de las cuales 193 serán polimorfismos genéticos.

Se ha conseguido tener un 8.04% de *missings* en la base de datos y un 7.16% de *missings* en las variables genéticas. Parece que los datos faltantes ya no ocasionarán tantos problemas, se cree que ya puede empezarse a trabajar.

VI. Estadística descriptiva y pruebas básicas

En este capítulo se presentan gráficos y datos descriptivos de las variables más importantes en cuatro grupos: variables demográficas, variables de rendimiento muscular, variables genéticas y otras variables.

6.1 Variables demográficas

Género

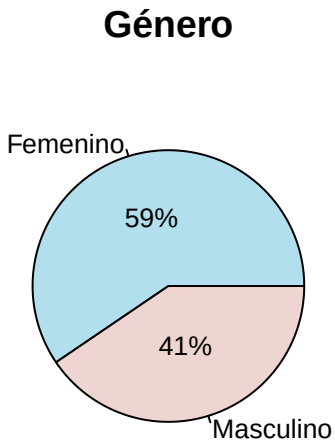

Figura 11: Gráfico de pastel del género.

Hay un 59% de mujeres y un 41% de hombres (Figura 11).

Edad

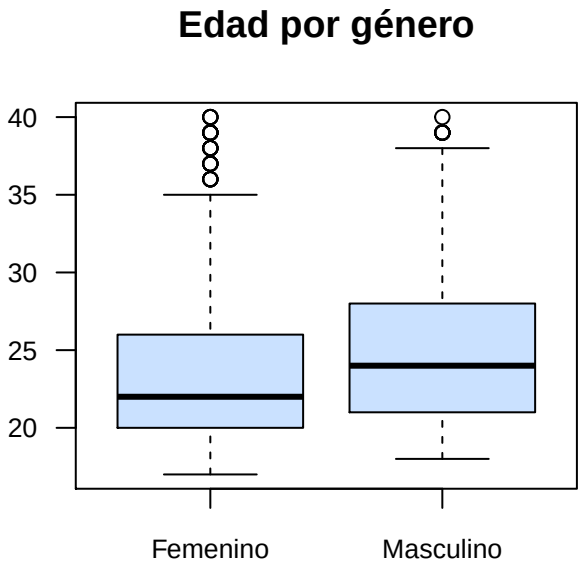

Figura 12: Boxplot de la edad estratificado por género.

La mayoría de las mujeres están entre los 20 y los 25 años aproximadamente, mientras que los hombres prensentan una edad que parece mayor. Se observa que, en general, son personas jóvenes, hecho que era de esperar ya que el estudio es llevado a cabo en estudiantes (Figura 12).

Raza

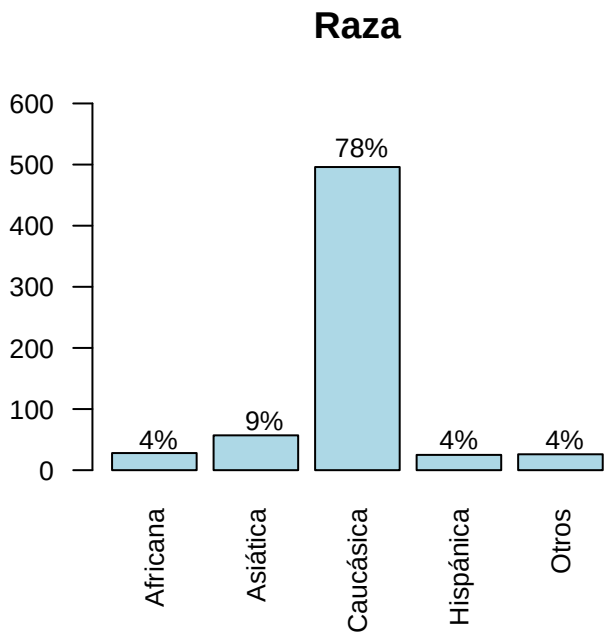

Figura 13: Gráfico de barras de Raza.

La mayoría de los individuos estudiados son de raza caucásica. En menor medida se observan afroamericanos, hispánicos, asiáticos y el grupo de otros (Figura 13).

Mano dominante

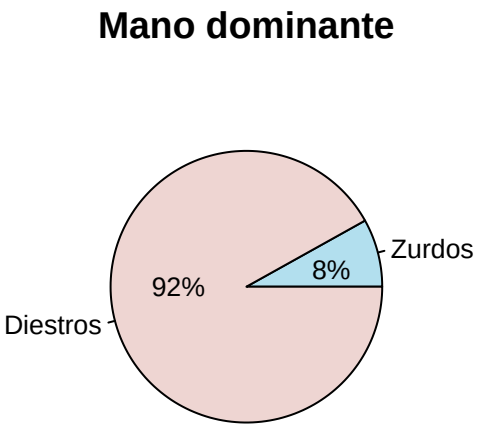

Figura 14: Gráfico de pastel de brazo dominante.

Hay un porcentaje de diestros muy superior al de zurdos (Figura 14).

6.2 Variables de rendimiento muscular

A continuación, se mostrarán dos variables que se consideran importantes referidas a secciones transversales de bíceps y tríceps y las variables respuesta de los modelos referidas a la cantidad de fuerza. En las cuatro variables pretende averiguarse si hay diferencias entre antes y después del entrenamiento, en otras palabras, si este ha surtido efecto.

Sección transversal bíceps

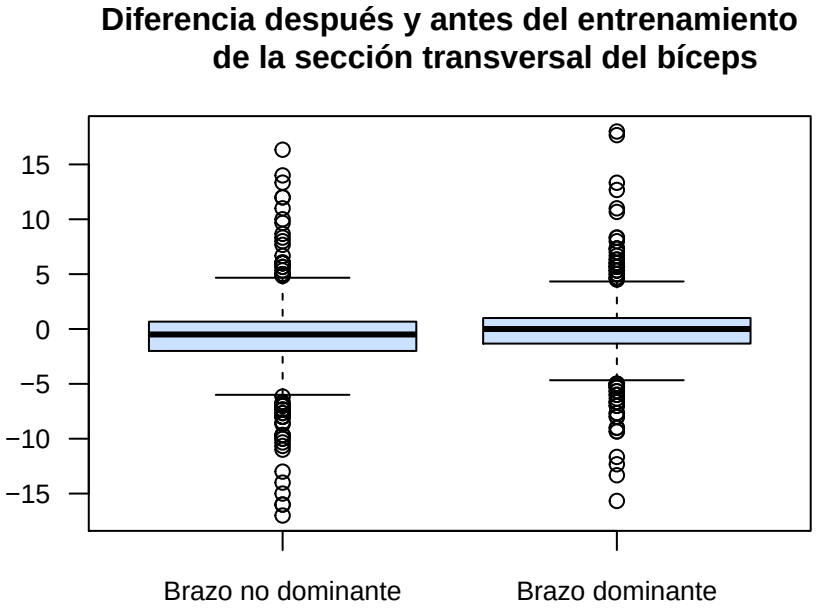

Figura 15: Boxplot de la ganancia de la sección transversal del bíceps estratificado por brazo dominante y no dominante.

Para el brazo dominante, la mediana parece estar centrada en el 0 y tener una distribución bastante simétrica. Además, no parece haber un aumento en el tamaño de la sección transversal del bíceps. En el caso del brazo dominante, la distribución también parece ser simétrica, sin embargo, la mediana parece un poco superior (Figura 15).

Se pretende comparar si la media de los dos grupos es la misma, para ver si existen diferencias entre grupos. Se utilizará la función `t.test` de la versión 3.4.3 del paquete `stats` de R para realizar una prueba t de Student para ver si las medias de los dos grupos son o no iguales.

$$\begin{cases} H_0 : \mu_{ND} = \mu_D \\ H_1 : \mu_{ND} \neq \mu_D \end{cases}$$

Esta función necesita que se especifique si las varianzas de los dos grupos son también o no iguales, por lo que primero se utilizará la función `var.test` de la versión 3.4.3 del paquete `stats` de R que computa una prueba F de Fisher cuyas hipótesis son:

$$\begin{cases} H_0 : \sigma^2_{ND} = \sigma^2_D \\ H_1 : \sigma^2_{ND} \neq \sigma^2_D \end{cases}$$

```
    F test to compare two variances

data:  bicND and bicD
F = 1.2082, num df = 626, denom df = 626, p-value = 0.01809
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.032876 1.413369
sample estimates:
ratio of variances
          1.208236
```

El estadístico de contraste es 1.208, cuyo p-valor asociado es 0.01809. Fijando un nivel de significación de 0.05 se tienen evidencias estadísticas suficientes para rechazar H_0 y decir que las varianzas son distintas. A continuación, se realizará el test de comparación de medias con 626 grados de libertad sabiendo que las varianzas son distintas. Se realizará un test aparejado, ya que las medidas se realizan en los dos brazos para cada persona.

```
    Paired t-test

data:  bicND and bicD
t = -5.2357, df = 626, p-value = 2.247e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.7199703 -0.3272068
sample estimates:
mean of the differences
             -0.5235885
```

El estadístico de contraste t es -5.2357, cosa que indica que la media del segundo grupo es superior a la del primero, la media del brazo dominante es superior a la del no dominante. Además, esta diferencia en tamaños para cada uno de los brazos es significativa, ya que el p-valor asociado al estadístico es 2.247e-07, extremadamente pequeño y, por supuesto, inferior al valor crítico fijado de 0.05. Se tienen evidencias estadísticas suficientes para rechazar la hipótesis nula y decir que existen diferencias entre las medias de los grupos. El tamaño de la sección transversal de bíceps ha aumentado más en el brazo dominante, no entrenado, que no dominante, entrenado. Este fenómeno puede asociarse a que la musculación hace que se reduzca la cantidad de grasa del músculo y por lo tanto su tamaño (Stewart y Rittweger, 2006).

Sección transversal tríceps

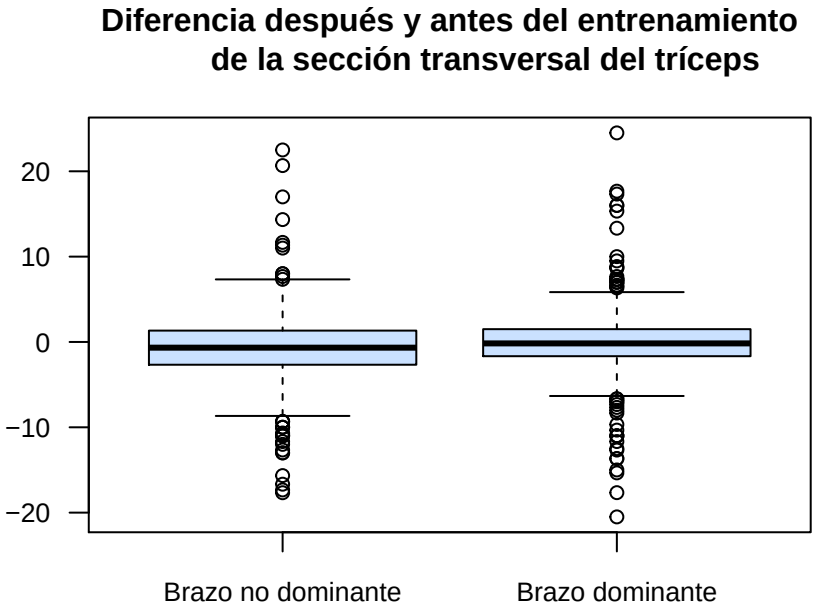

Figura 16: Boxplot de la ganancia de la sección transversal del tríceps estratificado por brazo dominante y no dominante.

Para el brazo dominante la mediana parece estar centrada en el 0 y tener una distribución bastante simétrica. No parece que aumente el tamaño de la sección transversal del bíceps. El caso del brazo dominante parece tener también una distribución simétrica, sin embargo, la mediana parece un poco superior (Figura 16).

Igual que para el bíceps, se pretende comparar si la media de los dos grupos es la misma, para ver si existen diferencias entre grupos. Se vuelve utilizar la función `t.test` de la versión 3.4.3 del paquete `stats` de R para realizar una prueba t de Student para ver si las medias de los dos grupos son o no iguales.

$$\begin{cases} H_0 : \mu_{ND} = \mu_D \\ H_1 : \mu_{ND} \neq \mu_D \end{cases}$$

Se vuelve a utilizar la función `var.test` de la versión 3.4.3 del paquete `stats` de R que computa una prueba F de Fisher para contrastar si las varianzas de los dos grupos son o no iguales cuyas hipótesis son:

$$\begin{cases} H_0 : \sigma^2_{ND} = \sigma^2_D \\ H_1 : \sigma^2_{ND} \neq \sigma^2_D \end{cases}$$

```
    F test to compare two variances

data:  triND and triD
F = 1.105, num df = 626, denom df = 626, p-value = 0.2117
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9446631 1.2926602
sample estimates:
ratio of variances
          1.105047
```

El estadístico de contraste es 1.105, cuyo p-valor asociado es 0.2117. Fijando un nivel de significación de 0.05 se concluye que no se tienen evidencias estadísticas suficientes para rechazar H_0 y decir que las varianzas son distintas. A continuación, se realizará el test de comparación de medias con 626 grados de libertad asumiendo que las varianzas no son distintas. Igual que anteriormente, se realizará un test aparejado ya que las medidas se realizan en los dos brazos para cada persona.

```
    Paired t-test

data:  triND and triD
t = -5.9241, df = 626, p-value = 5.182e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.9207414 -0.4622889
sample estimates:
mean of the differences
             -0.6915152
```

El estadístico de contraste t es -5.924, cosa que indica que la media del la ganancia de tamaño en la sección transversal del tríceps del brazo dominante es superior a la del no dominante. Además, esta diferencia en tamaños para cada uno de los brazos es significativa, ya que el p-valor asociado al estadístico es 5.182e-09, extremadamente pequeño y, por supuesto, inferior al valor crítico fijado de 0.05. Por lo que se tienen evidencias estadísticas suficientes para rechazar H_0 y decir que existen diferencias las medias de los grupos. El tamaño de la sección transversal de tríceps ha aumentado más en el brazo dominante, no entrenado, que no dominante, entrenado. Este fenómeno puede asociarse de la misma manera que para el bíceps a que la musculación hace que se reduzca la cantidad de grasa del músculo y por lo tanto su tamaño (Stewart y Rittweger, 2006).

Test de fuerza isométrica

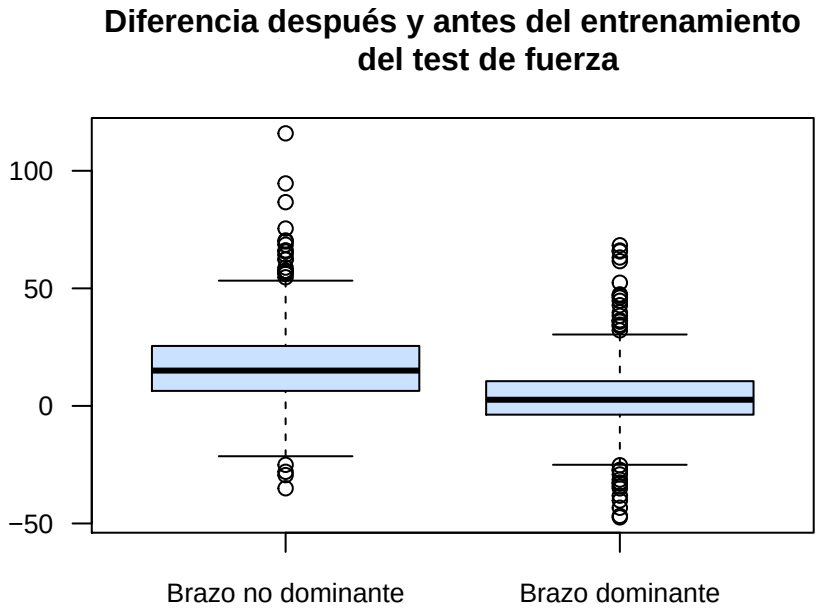

Figura 17: Boxplot de la ganancia en el test isométrico de fuerza estratificado por brazo dominante y no dominante.

En el brazo no dominante se puede observar que la mayoría de valores son positivos. Sin embargo, es importante destacar que hay algunos valores negativos de individuos que han perdido fuerza después de 12

semanas de entrenamiento. Por último, hay un grupo de personas que han ganado más fuerza que el resto (Figura 17).

En el brazo dominante puede observarse una distribución simétrica y centrada en el 0, ya que este brazo no ha recibido ningún tipo de entrenamiento.

En cuanto a las diferencias entre antes y después del entrenamiento, se compararan las medias de los dos grupos mediante la función `t.test` de la versión 3.4.3 del paquete `stats` de R como se ha hecho para las dos variables anteriores.

$$\begin{cases} H_0 : \mu_{ND} = \mu_D \\ H_1 : \mu_{ND} \neq \mu_D \end{cases}$$

Antes debe averiguarse si las varianzas muestrales pueden considerarse o no iguales mediante la función `var.test` de la versión 3.4.3 del paquete `stats` de R cuyas hipótesis son:

$$\begin{cases} H_0 : \sigma^2_{ND} = \sigma^2_D \\ H_1 : \sigma^2_{ND} \neq \sigma^2_D \end{cases}$$

```
    F test to compare two variances

data:  ND23_DIFF and D23_DIFF
F = 1.394, num df = 610, denom df = 606, p-value = 4.429e-05
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.188945 1.634463
sample estimates:
ratio of variances
          1.394048
```

El estadístico de contraste es 1.394, cuyo p-valor asociado es 4.429e-05. Se fija el nivel de significación en 0.05 como se ha hecho anteriormente. Se tienen evidencias estadísticas significativas suficientes para rechazar H_0 y decir que las varianzas son distintas. Se realizará el test T de Student con 606 grados de libertad sabiendo que las varianzas son distintas. Se realiza un test aparejado igual que en los otros apartados.

```
    Paired t-test

data:  ND23_DIFF and D23_DIFF
t = 25.446, df = 606, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 12.03626 14.04951
sample estimates:
mean of the differences
               13.04288
```

El estadístico de contraste t es 25.446, cosa que indica que la media del primer grupo es superior a la del segundo, la media del brazo no dominante es superior a la del dominante. Esta diferencia en la ganancia de fuerza de ambos grupos es significativa, ya el p-valor es de casi 0. El entrenamiento ha hecho que los individuos ganen fuerza en el test isométrico de bíceps. Este hecho concuerda con lo que se veía en el gráfico realizado (Figura 17)

Test de repetición máxima

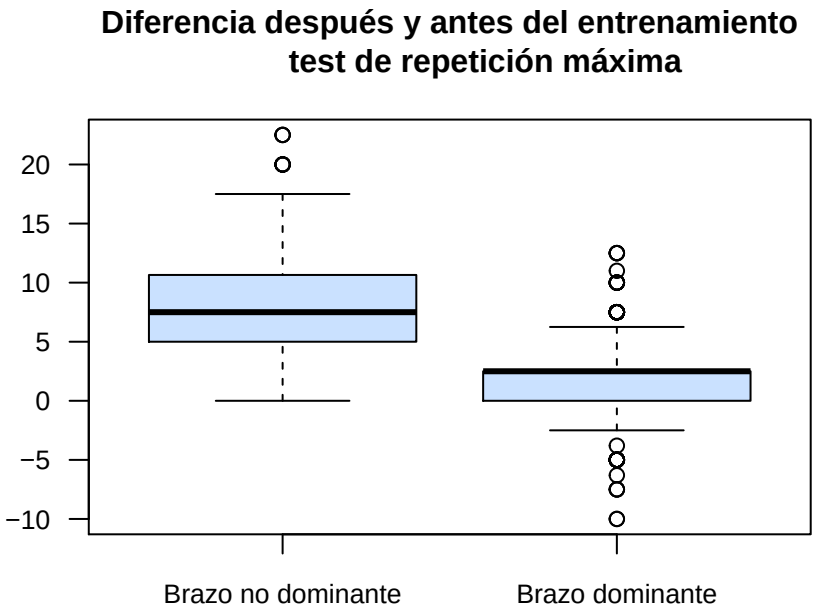

Figura 18: Boxplot del test de repetición máxima estratificado por brazo dominante y no dominante.

Estos *boxplots* son parecidos a los del apartado anterior. Puede observarse que la ganancia de fuerza es muy superior para el brazo no dominante, que ha sido entrenado, a la del brazo dominante, que no lo ha sido. Incluso parece que hay más diferencia entre un brazo y el otro (Figura 18.

Mediante el test T de Student se compararán las medias para ver si hay diferencias entre antes y después del entrenamiento, las hipótesis son:

$$\begin{cases} H_0 : \mu_{ND} = \mu_D \\ H_1 : \mu_{ND} \neq \mu_D \end{cases}$$

Para averiguar si las varianzas muestrales pueden considerarse o no iguales se realiza el test F de Fisher con las hipótesis:

$$\begin{cases} H_0 : \sigma^2_{ND} = \sigma^2_D \\ H_1 : \sigma^2_{ND} \neq \sigma^2_D \end{cases}$$

```
    F test to compare two variances

data:  NDRM_DIFF and DRM_DIFF
F = 1.6719, num df = 618, denom df = 621, p-value = 2.016e-10
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.428074 1.957369
sample estimates:
ratio of variances
           1.67188
```

El estadístico de contraste es 1.672, el p-valor asociado es 2.016e-10. Con el mismo nivel de significación que en los otros apartados, se tienen evidencias estadísticas significativas para rechazar H_0 y decir que las

varianzas son distintas. Se realizará el test T de Student con 613 grados de libertad sabiendo que las varianzas son distintas. Se vuelve a realizar un test aparejado.

```
    Paired t-test

data:  NDRM_DIFF and DRM_DIFF
t = 40.902, df = 613, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 6.583129 7.247164
sample estimates:
mean of the differences
               6.915147
```

El estadístico de contraste t es 40.902, cosa que indica que la media del primer grupo es superior a la del segundo, la media del brazo no dominante es superior a la del dominante. Esta diferencia en la ganancia de fuerza en el test de repetición máxima de ambos grupos es significativa, ya el p-valor es extremadamente pequeño.

Según los resultados obtenidos en este apartado parece que en las 4 variables más importantes de rendimiento muscular se observa diferencia estadística significativa antes y después del estudio, se observan diferencias en el brazo entrenado y no entrenado. Parece que el entrenamiento ha surtido efecto.

Comparación test de ganancia de fuerza isométrica vs repetición máxima

Se cree que existe una correlación entre los individuos que han ganado más fuerza en el test de fuerza isométrica también la han ganado en el test de repetición máxima. Se realizarán 5 gráficos (Figura 19), uno para cada raza, en los que cada individuo será representado por una línea. En cada gráfico se representarán las dos variables de ganancia de fuerza, cosa que permitirá observar gráficamente si es cierto que los individuos que han ganado más fuerza en un test también lo han hecho en el otro. Se espera que las rectas que unen un mismo individuo sean más o menos paralelas y que no haya muchos cruces.

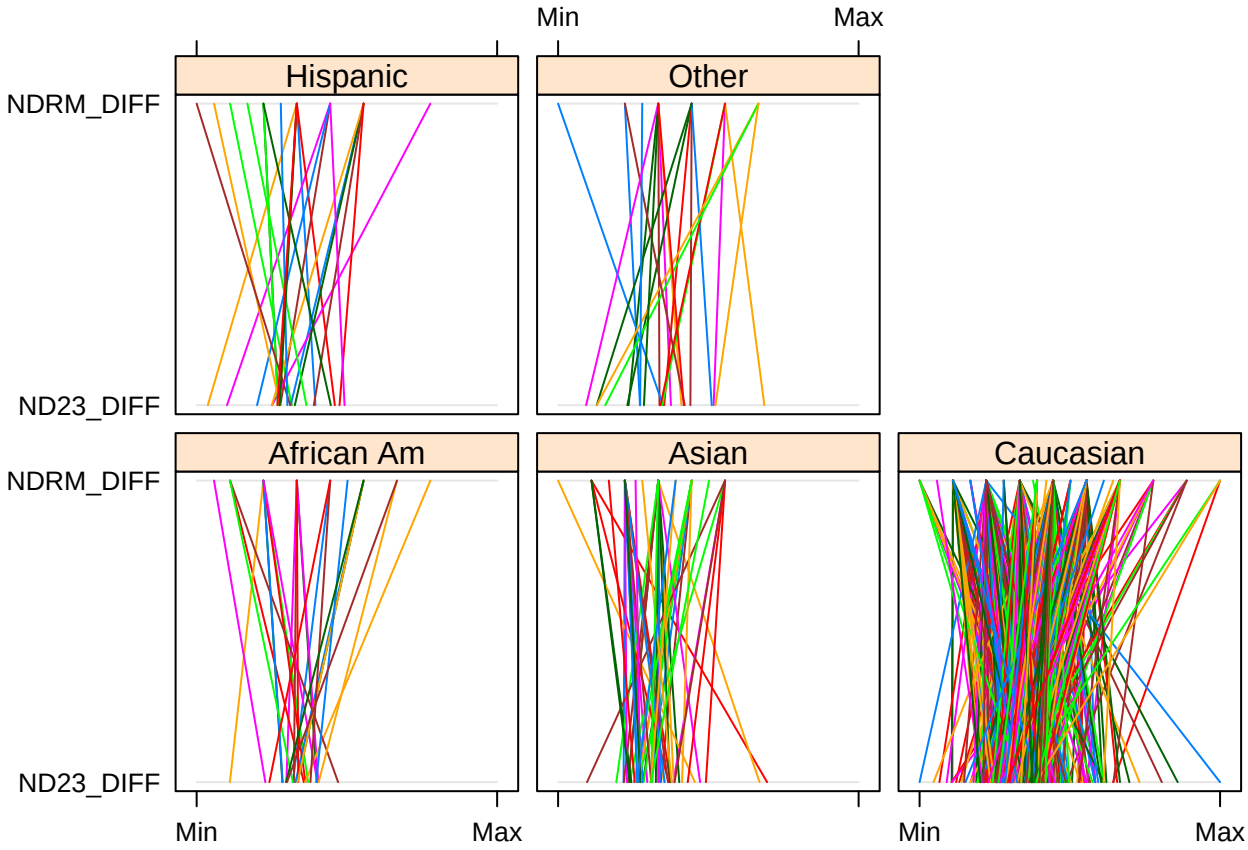

Figura 19: Gráfico comparativo intra-individuos de la ganancia en el test de repetición máxima y de fuerza isométrica.

El primer aspecto a destacar es que, como ya se sabía, hay muchos más individuos de raza caucásica, cosa que no permite concluir nada de esta raza. Contrariamente, en las otras 4 razas sí que se nos permite observar que la mayoría de las rectas se cruzan, por lo que parece que no se cumple la hipótesis y que las personas que ganan más fuerza en el test de fuerza isométrica no tienen por qué ganar más fuerza en el test de repetición máxima. Por otro lado, parece que hay más variabilidad en la prueba de fuerza isométrica que en la de repetición máxima.

6.3 Variables genéticas

Frecuencia del alelo menos común

La frecuencia del alelo menos común (MAF, por sus siglas en inglés) es la frecuencia del alelo que tiene la menor proporción de todos en una población dada. Se utiliza para diferenciar entre variantes comunes y raras dentro de una población. Interesa estudiar SNPs con una MAF alta, cercana a 0.5, ya que son más diversos.

Se ha calculado y graficado la frecuencia del alelo menos común de cada SNP para cada raza y para el total (Figura 20).

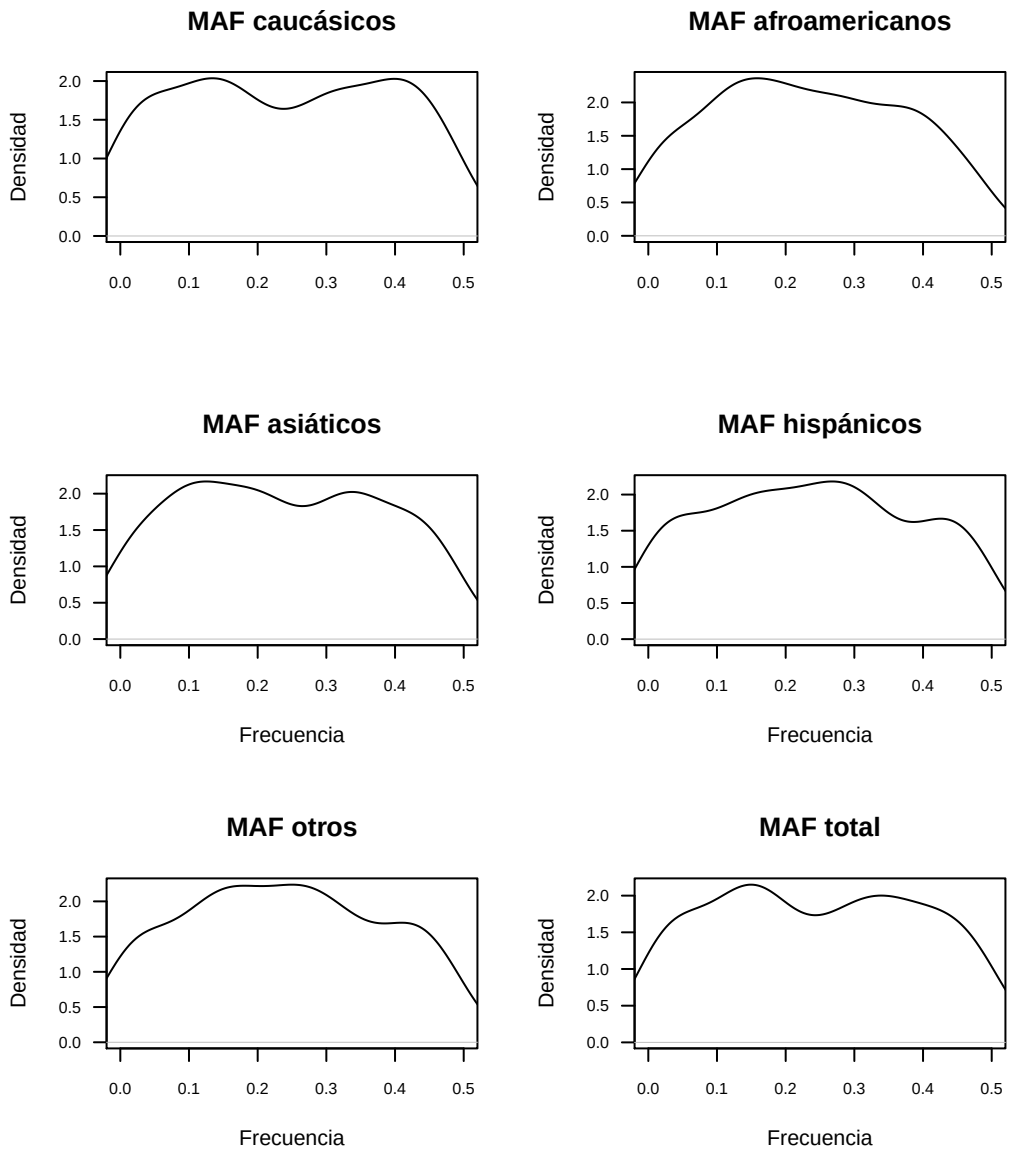

Figura 20: Gráfico de densidad de la frecuencia del alelo menos común por razas y para el total.

En los gráficos anteriores puede observarse que para todos los grupos el centro de la distribución parece estar entre 0.20 y 0.30, por lo que se puede decir que en general los polimorfismos tienen una frecuencia substancial de ambos alelos.

Ley de Hardy-Weinberg

El equilibrio de Hardy-Weinberg establece que la composición genética de una población permanece en equilibrio de generación en generación mientras no se produzca ninguna mutación ni actúe la selección natural ni ningún otro factor; ya que la herencia por sí misma no engendra ningún cambio evolutivo.

El principio de Hardy-Weinberg también establece que las frecuencias de los genotipos de un *locus* individual se fijarán en un equilibrio particular, bajo ciertas condiciones y tras una generación de apareamiento al azar.

En general, en los estudios de asociación genética, se eliminan SNPs muy significativos respecto a la prueba de equilibrio, por la sospecha de errores de genotipado (confusión entre homocigotos y heterocigotos)

Raza	Porcentaje significativo
Caucásicos	12.95%
Afroamericana	3.11%
Asiática	8.29%
Hispánica	4.66%
Otras	4.15%
Total	24.35%

Tabla 1: Tabla porcentaje significativo en el test de Hardy-Weinberg por raza

Se utiliza el test `HWExact` de la versión 1.6.1 del paquete `HardyWeiberg` (Jan Graffelman) que realiza el test exacto para el equilibrio de Hardy-Weinberg. Se calcula el porcentaje significativo de SNPs por cada una las razas y el total. Dicho porcentaje no cumplirá el equilibrio de Hardy-Weinberg.

Las hipótesis para las frecuencias alélicas que contrasta este test son:

$$\begin{cases} H_0 : 0 \to \ p^2, \ 1 \to 2pq \ \ y \ \ 2 \to q^2 \\ H_1 : de\ otra\ manera \end{cases}$$

Porcentaje significativo por raza (valor crítico 0.05)

Las razas afroamericana, hispánica y el grupo de "otros" tienen un porcentaje de SNPs que no cumplen el equilibrio de Hardy-Weinberg pequeño, entre 3 y 4 % (Tabla 1). Sin embargo, las razas caucásica y asiática, tienen un porcentaje que de 12.95% y 8.29% respectivamente, se estudiará si alguno de estos SNPs debería excluirse de la base de datos. Se deberá tener en cuenta que debido a la gran cantidad de pruebas de significación que se hacen, es normal que aparezcan alrededor de un 5% de falsos positivos.

Por otro lado, el porcentaje significativo del todos los datos, sin tener en cuenta las distintas razas, es muy superior (ronda el 24%). Este hecho no sorprende ya que es sabido que las poblaciones alélicas son heterogéneas, por lo que se dispara el desequilibrio.

Los p-valores de cada uno de los grupos deberían seguir a una distribución uniforme. Se graficará un Q-Q plot con $-log_{10}(pval)$ con dicha distribución para comprobar si distan de la misma. Se utilizará la función `qqunif` de la versión 1.1-22 del paquete `gap` (Figura 21).

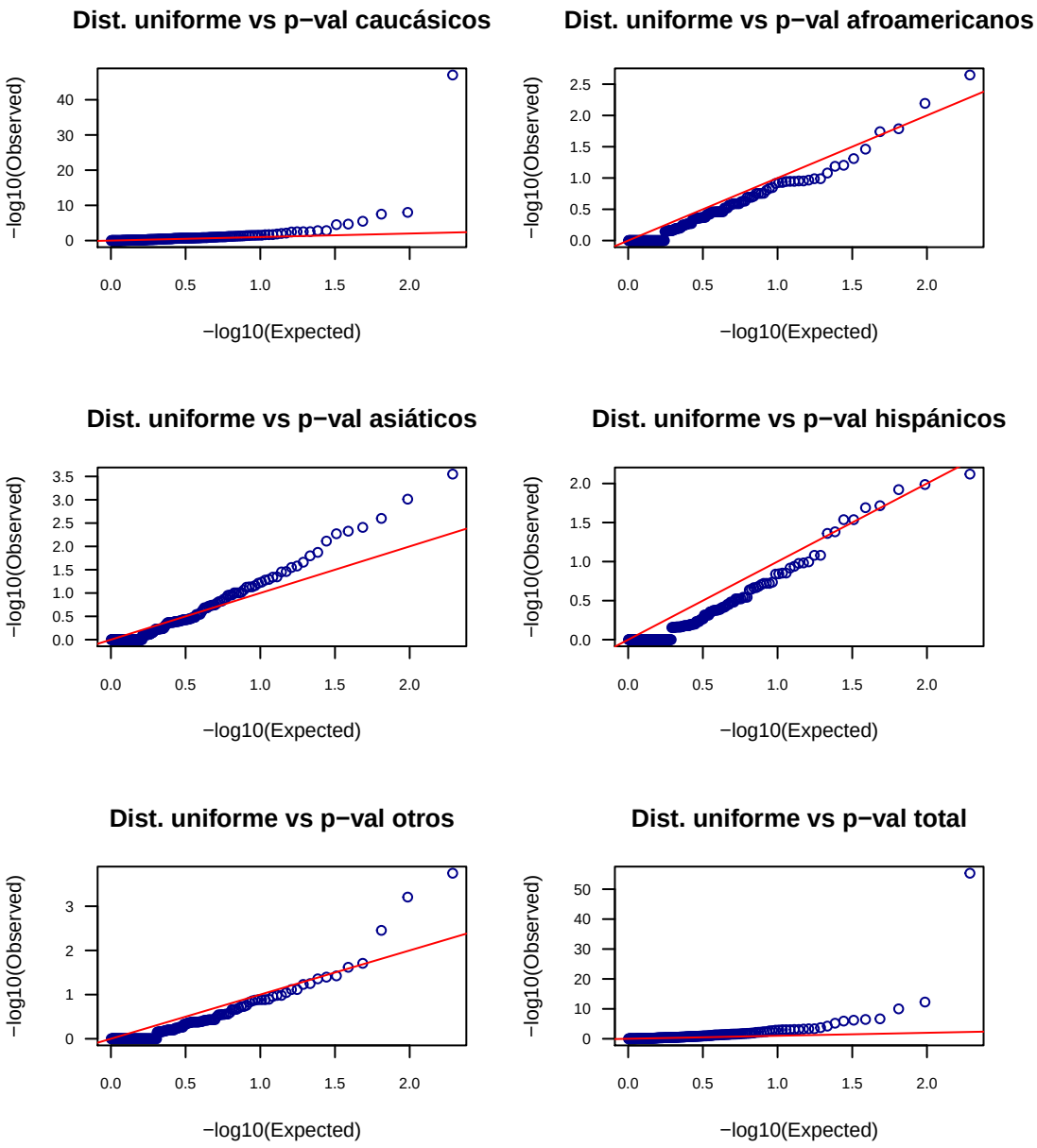

Figura 21: Q-Q plot de los p-valores del test de Hardy-Weinberg por raza.

Los afroamericanos y los hispánicos si que parece que se adapten bien a la distribución uniforme. Sin embargo, en el resto se puede observarse que la cola de la izquierda sí que parece adecuarse a una distribución uniforme mientras que los últimos valores parecen ser p-valores muy pequeños. Parece que se podría dudar de la calidad del marcador más significativo de los caucásicos. En el gráfico del total puede observarse el mismo fenómeno (Figura 21).

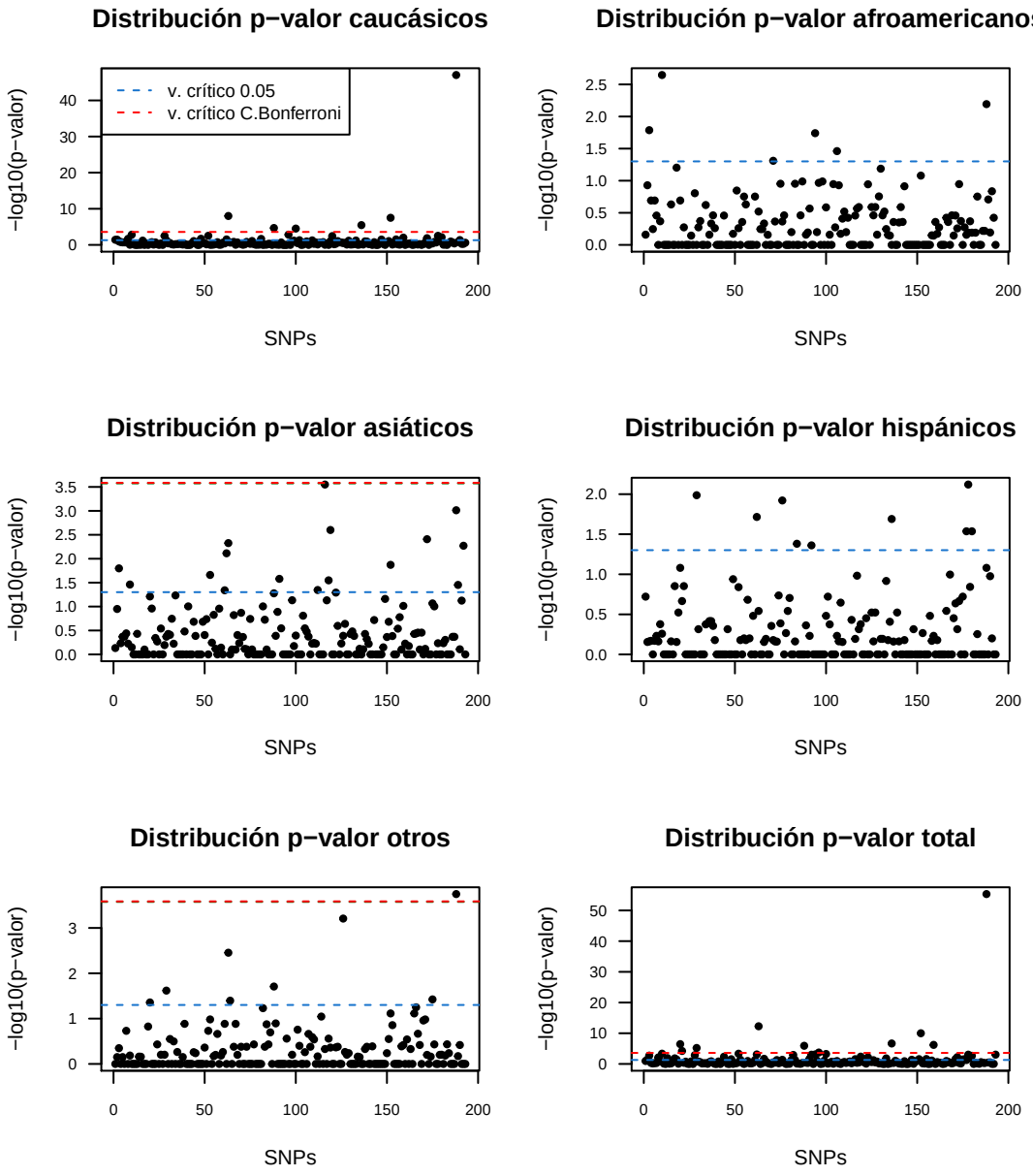

Figura 22: Distribución de los p-valores del test de Hardy-Weinberg por raza y para el total.

Esta representación gráfica de cada una de las razas y el total permite ver qué y cuántos SNPs están por encima de los valores fijados como crítico (Figura 22). También se ha graficado el valor crítico con la Corrección de Bonferroni para evitar el problema de la multiplicidad, sin embargo, es un test muy conservador. Esta corrección establece el valor crítico en $\alpha_{Bonferroni} = \frac{0.05}{193} = 0.0002591$.

Puede observarse que en todos los grupos hay SNPs por encima del valor crítico 0.05, sin embargo, como habíamos visto anteriormente sólo en los individuos afroamericanos, en los del grupo otros y en el total hay SNPs significativos con la Corrección de Bonferroni.

Se propone otro método alternativo no tan conservador para controlar los falsos positivos: el *False Discovery Rate* (FDR). Este método se basa en la proporción esperada de falsos positivos de entre todos los test considerados como significativos. Se utilizará la función de la versión 1.9 del paquete `astsa` que computa este método con la aproximación de Benjamini & Hochberg con un máximo del 5% de resultados estadísticamente significativos. Según este método hay 8 SNP significativos en los caucásicos, uno en el grupo de otros y 25

para el total, que representan el 4.1%, 0.5% y 12% de los polimorfismos.

Se espera alrededor de 9-10 SNPs significativos en cada raza (5% de 193) a causa de los falsos positivos por la gran cantidad de test que se realizan. Puede observarse que este valor es superior en algunos grupos e inferior en otros, no parece que ocurra algo no esperado. Sin embargo, en el grupo de los caucásicos sí que se observa un SNP con un p-valor extremadamente pequeño (5.8604e-57), casi 0. Además, esta raza es la que tiene más individuos, por lo que este valor es más representativo. Dicho SNP, "ucp2_ct", se eliminará de la base de datos para todos los individuos.

6.4 Otras

Centro

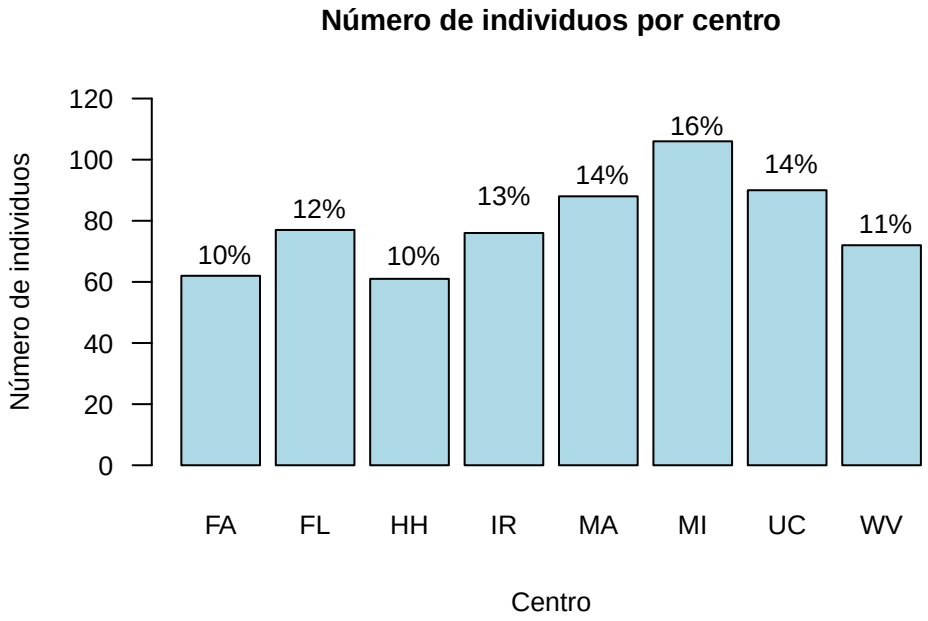

Figura 23: Diagrama de barras de la frecuencia de los individuos en los centros de estudio.

Los datos se recogieron en 8 centros, con un número de observaciones parecido (Figura 23). Se tiene: "FA" para *Florida Atlantic University*, "FL" para *University of Central Florida*, "HH" para Hartfort Hospital, "IR" para *Dublin City University* (Irlanda), "MA" para *University of Massachusetts*, "MI" para *Central Michigan University*, "UC" para *University of Connecticut* y "WV" para *University of West Virginia*.

Trimestre

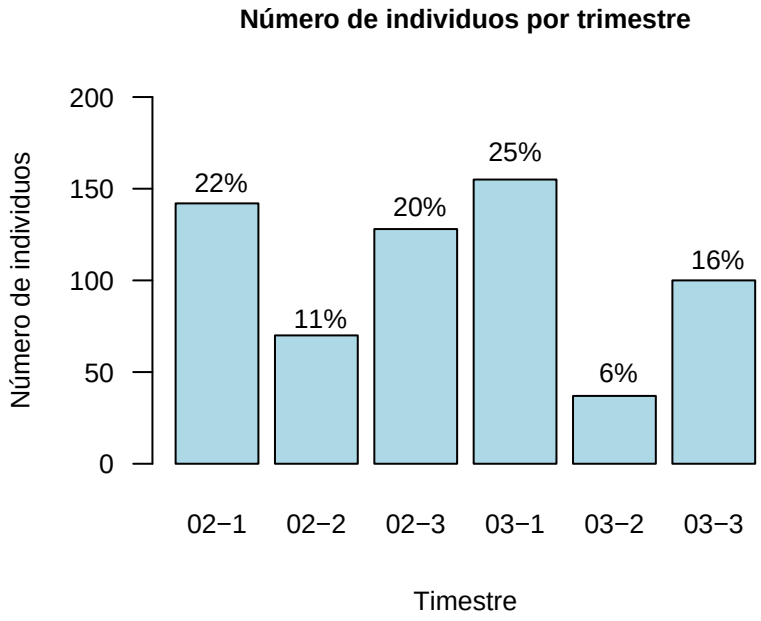

Figura 24: Diagrama de barras de la frecuencia de los individuos en cada trimestre.

En los segundos trimestres de 2002 y 2003 parece que el número de individuos estudiados es menor (Figura 24).

VII. Exploración de los datos genéticos

7.1 Preparación

Recodificación de los datos genéticos

Se recodifican los datos genéticos como *2* para el homocigoto menos frecuente, 0 para el más frecuente y 1 para los heterocigotos.

Hay un 7.12% de *missings* por variable y por individuo de los datos que faltan. Se representa gráficamente el porcentaje de *missings* por SNP y la media de datos faltantes (Figura 25).

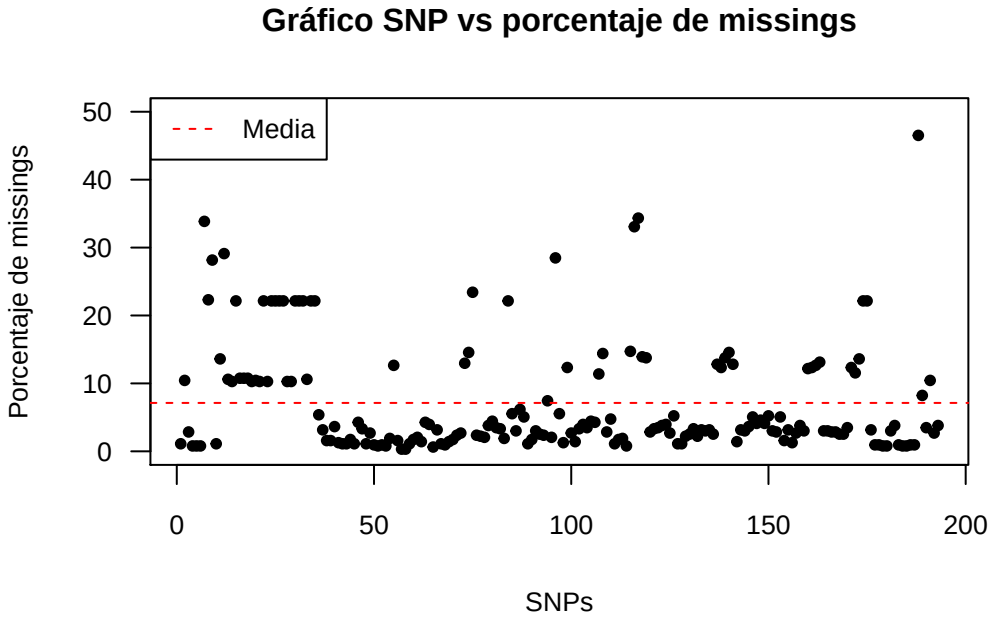

Figura 25: Gráfico SNP vs porcentaje de missings y media.

Puede observarse que hay cuatro SNPs con un porcentaje de datos faltantes que rondan el 30-50%. Sin embargo, la mayoría están entre el 0 y 10%.

7.2 Imputación de missings

Se considera que los datos faltantes son *Missing Completely Random* (MCAR), completamente aleatorios, que no existe relación entre si falta un dato y cualquier otro valor en el conjunto de datos, faltado u observado. Por lo tanto, los datos que faltan son solamente un subconjunto aleatorio de los datos completos. La imputación de *missings* consiste en sustituir estos valores por un valor de la base de datos. Para algunas posteriores técnicas de análisis se necesita una base de datos sin valores faltantes. Mediante la distribución trinomial se calcula la proporción en la que aparece cada uno de los niveles del factor para cada SNP. Con la función `sample` (versión 3.4.3 del paquete `base`) se obtiene una muestra aleatoria del número de *missings* que haya en el SNP con las probabilidades obtenidas y se sustituyen por los valores NA. Al sustituir los *missings* por valores aleatorios de la distribución trinomial, la base de datos ya no tiene valores faltantes en las variables genéticas.

7.3 Eliminación monomórficos

Un SNP monomórfico hace referencia a la existencia en una sola forma alélica de un gen. Se eliminan los 11 SNPs monomórficos ya que al no tener variabilidad no aportan ningún tipo de información para el estudio.

7.4 Análisis de componentes principales

Se ha utilizado el análisis de componentes principales (ACP o PCA) para analizar las variables genéticas de manera visual. Se sospecha que se tiene una población no homogénea a causa de las distintas etnias incluidas en el estudio.

Este tipo de análisis organiza los datos en "nuevas variables", llamadas componentes, que se ordenan por la cantidad de varianza de los datos originales que recogen. Estas variables no están correlacionadas. Es una técnica útil para reducir la dimensionalidad de los datos y poder representarlos gráficamente, cosa que cuando se tienen muchas variables es imposible. Esta técnica busca la proyección de los datos según la cual los datos quedan mejor representados en términos de mínimos cuadrados mediante el cálculo de la descomposición en autovalores de la matriz de covarianza.

Se utiliza la función `princomp` de la versión 3.4.3 del paquete `stats` de R para realizar el análisis. El cálcul se realizará con la matriz de varianzas y covarianzas para minimizar la pérdida de información útil posible, también podría hacerse con la matriz de correlaciones. El tamaño de la matriz de datos es de 645×193.

Se representa en el gráfico a continuación el porcentaje de la varianza de los datos originales recogida en cada componente (Figura 26).

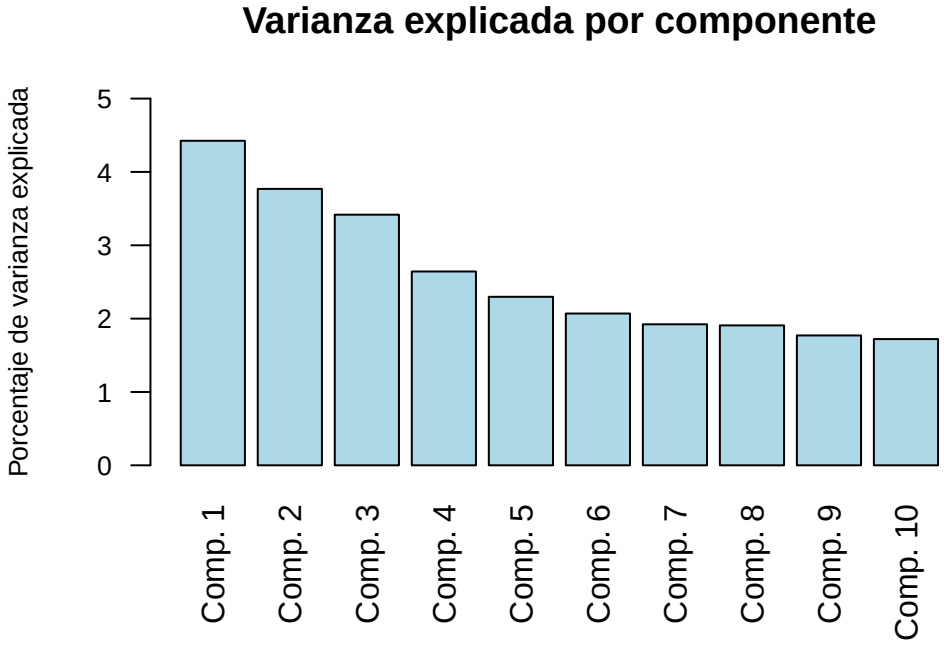

Figura 26: Varianza que explica cada componente en el análisis de componentes principales.

Entre la primera y la segunda dimensión se tiene casi 8% de varianza explicada, es una proporción pequeña, no obstante este hecho es habitual en este tipo de datos. Entre la segunda y la tercera componente se recoge un 7% de varianza mientras que con la primera y la tercera un 7.7%.

A continuación, se mostrarán los individuos según su raza en estas dimensiones.

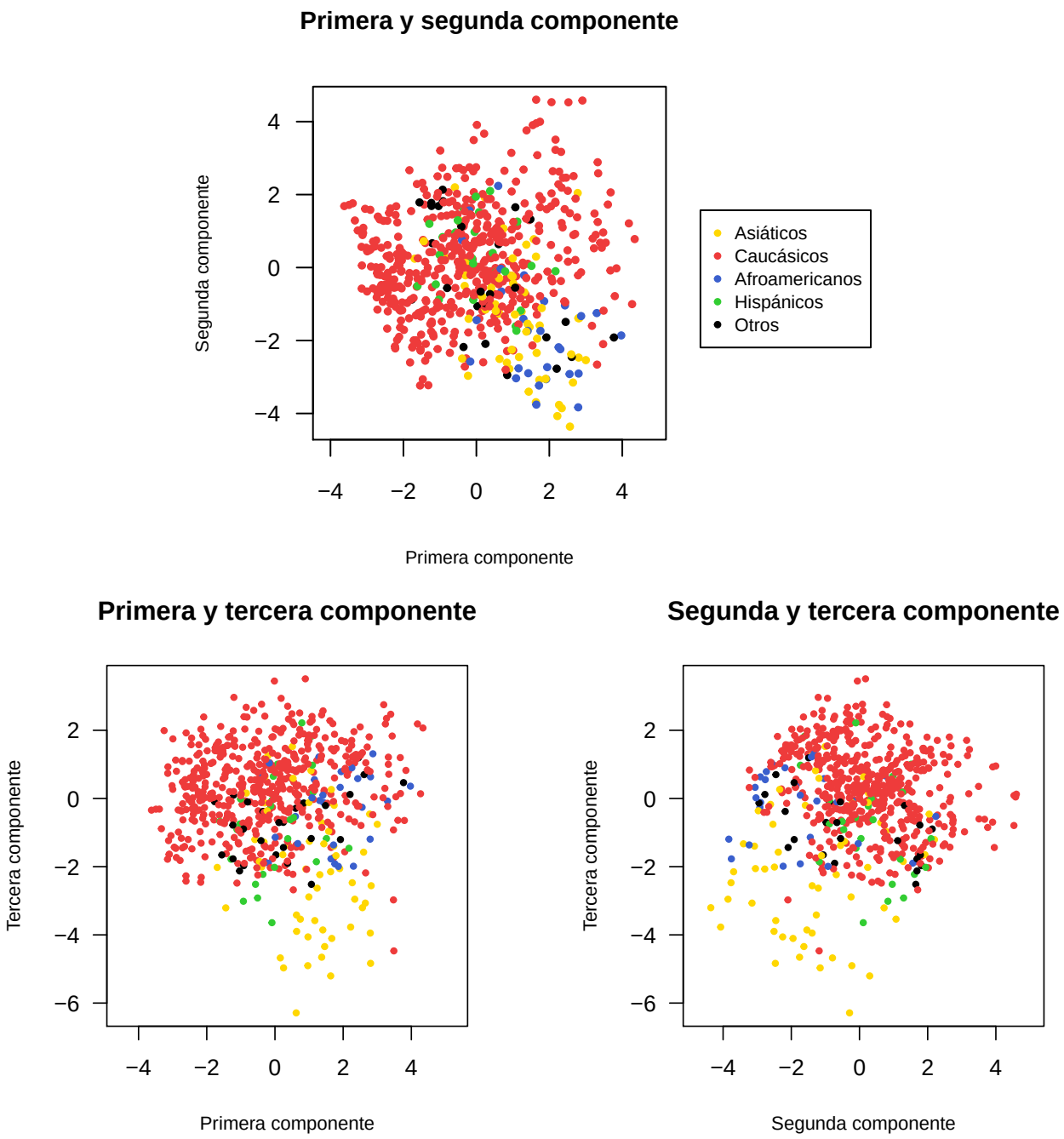

Figura 27: Proyecciones de los individuos en primera, segunda y tercera componentes del análisis de componentes principales.

En los tres gráficos parece que sí que haya relación entre las variables genéticas y la raza de los sujetos ya que parecen que estén más o menos agrupado (Figura 27). Los individuos de raza caucásica, la más numerosa, y los del grupo "otros" parecen aportar más a la primera y segunda dimensión, los de etnia asiática a la segunda y tercera, los afroamericanos a la segunda.

VIII. Modelos estadísticos de rendimiento muscular

Se han escogido 2 variables de las 76 que había de rendimiento muscular. Se ha considerado que eran las más representativas para medir el rendimiento muscular que servirán como variable respuesta.

- `ND23_DIFF`: es la diferencia entre `Post_ND_avg` y `V23_ND_AVG`, promedio de las medidas de después del entrenamiento y de los días 2 y 3 antes del entrenamiento en el brazo no dominante. Podría decirse que es la fuerza que se ha ganado con el entrenamiento.
- `NDRM_DIFF`: es la diferencia del test de repetición máxima de bíceps entre después y antes del entrenamiento del brazo no dominante

Para cada una de estas variables se han considerado cuatro modelos, en cada uno la variable respuesta sigue un esquema distinto: aditivo, recesivo, dominante y codominante. A continuación, se mostrará, a modo de ejemplo, cada uno de estos esquemas en la Figura 28.

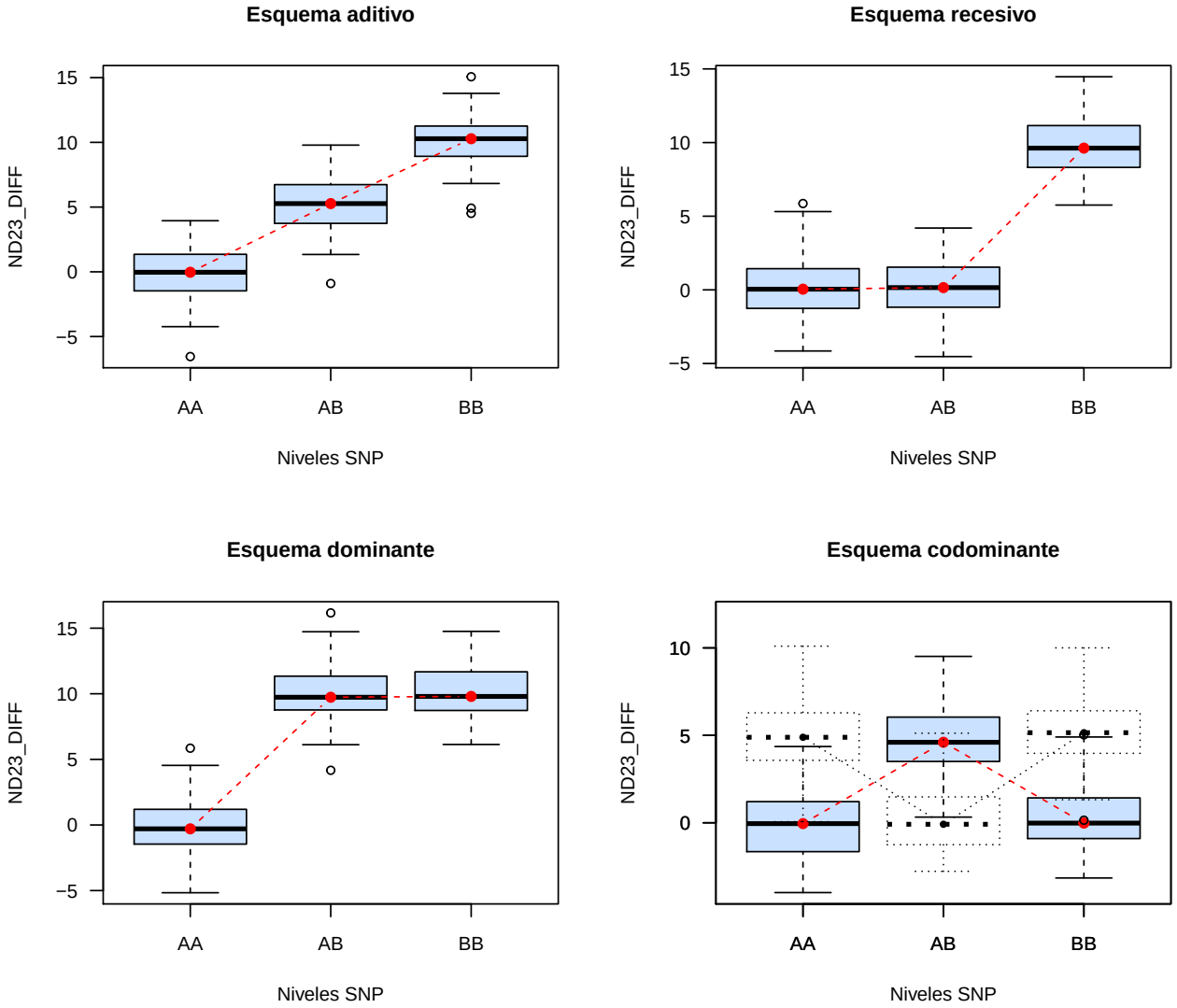

Figura 28: Ejemplos de esquema aditivo, recesivo, dominante y codominante.

En el modelo aditivo, se considera el efecto del homocigoto menos frecuente como el que más efecto tiene en la variable respuesta (en este caso "BB"), seguido del heterocigoto ("AB") y, por último, el homocigoto más

frecuente ("AA"). Se mantiene una relación lineal entre el número de alelos "B" y la variable respuesta. En este modelo se considera el polimorfismo como variable numérica, los componentes se agregan de manera conjunta para modelar los datos.

En el modelo recesivo, se supone que solamente los individuos que tienen dos copias del alelo menos frecuente ("BB" en este caso) tienen efecto en la variable respuesta, mientras que los que tienen una copia o ninguna del alelo ("AA" o "AB") no tienen efecto. El SNP se considera también como variable numérica.

En el caso del modelo dominante, se supone que los individuos que tienen, por lo menos, una copia del alelo menos frecuente ("AB" o "BB") tienen efecto en la variable respuesta. Los sujetos que no tienen ninguna copia de este alelo no tienen efecto sobre la variable respuesta. También se considera el polimorfismo como variable numérica.

Por último, el modelo codominante no tiene ningún supuesto exacto sobre los genes, es el más flexible y permite cualquier patrón. En este esquema, el polimorfismo se considera como variable cualitativa.

8.1 Ganancia de fuerza isométrica (`ND23_DIFF`)

Selección del modelo

En primer lugar, se considerará un modelo con la variable respuesta `ND23_DIFF` y las variables `Center`, `Term`, `Gender`, `Age`, `Race`, `Pre.weight`,`Pre.height`, `pre.BMI`, `SBP`, `DBP`, `FLGU`, `TG`, `CHOL`, `HDL_C`, `CHOL_HDL_C`, `VLDL_TG`, `LDL_C`, `FINS`, `CRP`, `HOMA`, `Met_syn` como explicativas.

A continuación, se utilizará la función `stepAIC` de la versión 7.3-48 del paquete `MASS` para realizar una selección de variables adecuada según el Criterio de Información de Akaike (AIC). Este Criterio es una medida de calidad relativa de un modelo estadístico. El AIC tiene en cuenta la bondad del ajuste del modelo y su complejidad. Permite comparar modelos anidados. El mejor modelo será el que tenga un AIC menor.

Se consideran las posibles interacciones de primer orden entre los polimorfismos, sin embargo, no se utilizarán en el modelo final ya que no son estadísticamente significativas, no ayudan a explicar la variación de fuerza de los individuos.

El test que contrasta si el coeficiente asociado al parámetro de cada variable es o no distinto de 0:

$$\begin{cases} H_0 : \beta = 0 \\ H_1 : \beta \neq 0 \end{cases}$$

Seguidamente, se computará el modelo y se mostrará, entre otras cosas, los resultados del test. En principio variables que tengan asociado un p-valor más pequeño que 0.05, fijado como valor crítico, se confirmará que se incluyen en el modelo, sin embargo, se tendrán en cuenta también otras pruebas que vendrán a continuación.

```
Call:
lm(formula = ND23_DIFF ~ Center + Term + Gender + DBP + VLDL_TG +
    Race, data = data2)

Residuals:
    Min      1Q  Median      3Q     Max
-49.328  -8.148  -0.854   7.729  80.628

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.35598    7.81752   0.813  0.41663
CenterFL     -2.93104    2.78759  -1.051  0.29363
CenterHH     -0.58753    3.37862  -0.174  0.86203
```

```
CenterIR      21.95882   15.73478   1.396  0.16355
CenterMA       7.69799    3.26816   2.355  0.01894 *
CenterMI      -6.40466    2.69715  -2.375  0.01800 *
CenterUC       3.45543    2.75249   1.255  0.21001
CenterWV       8.40153    2.90160   2.895  0.00397 **
Term02-2      -0.75884    2.98707  -0.254  0.79958
Term02-3       3.71532    2.30128   1.614  0.10715
Term03-1      -5.05458    2.07752  -2.433  0.01537 *
Term03-2      -3.19016    4.05243  -0.787  0.43158
Term03-3      -2.06861    2.20315  -0.939  0.34828
GenderMale     7.48876    1.61361   4.641 4.58e-06 ***
DBP            0.17274    0.08584   2.012  0.04480 *
VLDL_TG       -0.15921    0.07596  -2.096  0.03666 *
RaceAsian     -5.40282    4.06593  -1.329  0.18460
RaceCaucasian -1.98588    3.29869  -0.602  0.54747
RaceHispanic  -8.08848    4.45001  -1.818  0.06980 .
RaceOther      2.59318    4.91103   0.528  0.59774
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.28 on 439 degrees of freedom
  (173 observations deleted due to missingness)
Multiple R-squared:  0.1904,    Adjusted R-squared:  0.1553
F-statistic: 5.433 on 19 and 439 DF,  p-value: 5.315e-12
```

Se duda de si las variables `Center` y `Term` deben incluirse en el modelo, ya que solamente algunos niveles de las mismas tienen un coeficiente asociado al parámetro de cada nivel significativo. Se realizará una prueba F con la función `anova` (versión 3.4.3 del paquete `stats`) para comparar el modelo con cada una de las variables y sin ellas para comprobar si son, o no, distintos.

$$\begin{cases} H_0 : modelo\ nulo = modelo\ ampliado \\ H_1 : modelo\ nulo \neq modelo\ ampliado \end{cases}$$

```
Analysis of Variance Table

Model 1: ND23_DIFF ~ Center + Term + Gender + DBP + VLDL_TG + Race
Model 2: ND23_DIFF ~ Term + Gender + DBP + VLDL_TG + Race
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1    439 102452
2    446 114610 -7    -12158 7.442 1.807e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Analysis of Variance Table

Model 1: ND23_DIFF ~ Center + Term + Gender + DBP + VLDL_TG + Race
Model 2: ND23_DIFF ~ Center + Gender + DBP + VLDL_TG + Race
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    439 102452
2    444 106176 -5   -3723.1 3.1906 0.00768 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El p-valor asociado a las pruebas F son de $1.807e-08$ y 0.00768, en los dos casos por lo tanto, inferiores a 0.05, por lo que se rechaza la hipótesis nula de no efecto de `Center` y de no efecto de `Term` en el modelo.

Se concluye que los modelos son distintos y, por lo tanto, las variables `Center` y `Term` son importantes para explicar la variable respuesta. Deben incluirse en los modelos, cosa que en un principio no es buena ya que se supone que las pruebas que se realizan a los individuos están estandarizadas y son las mismas entre centros y trimestres. Parece que la estandarización no es del todo efectiva, y por lo tanto se controlará el efecto del centro y del trimestre en el modelo.

Se incluye la variable `Race` en el modelo aunque no sea estadísticamente significativa fijando 0.05 como valor crítico ya que se ha visto en el análisis de componentes principales y en el equilibrio de Hardy-Weinberg que las frecuencias alélicas son heterogéneas. Se cree que es importante la raza para una buena predicción del modelo.

Finalmente, el modelo que se utilizará es el que tiene como variable respuesta `ND23_DIFF` y como explicativas `Center` (centro en que se realizan las pruebas), `Term` (trimestre en que se realizan los test), `Gender` (género del individuo), `DBP` (presión sanguínea diastólica), `VLDL_TG` (lipoproteínas de muy baja densidad) y `Race` (raza del individuo).

El R^2_{adj} del modelo es 0.1553, por lo cual sabemos que el modelo explica el 15.53% de la variabilidad de la variable respuesta. El R^2_{adj} tiene en cuenta el número de variables explicativas del modelo. Recordamos que este modelo no incluye aún ninguna variable genética.

El estadístico F es un buen indicador de si existe, o no, relación entre los predictores y las variables respuesta. Cuanto más lejos esté el estadístico de 1 mejor será. Sin embargo, dicho estadístico también será más grande cuantas más variables explicativas se tengan y más pequeño cuantos más datos se tengan. Se contrasta si es igual que 1 teniendo en cuenta los grados de libertad que se tienen:

$$\begin{cases} H_0 : No\ hay\ relación\ entre\ la\ variable\ respuesta\ y\ las\ variables\ explicativas \\ H_1 : Hay\ relación\ entre\ la\ variable\ respuesta\ y\ las\ variables\ explicativas \end{cases}$$

El p-valor del contraste es de $5.315e-12$, por lo que se puede rechazar la hipótesis nula y decir que existe relación entre la variable respuesta y los predictores, es un modelo útil.

A continuación, se añadirá al modelo cada uno de los SNPs como variable explicativa con tal de poder identificar qué polimorfismos influyen más en la ganancia de fuerza en la prueba.

Como se ha dicho anteriormente, cada marcador genético tiene 3 niveles: 0, 1 y 2. Como los polimorfismos pueden seguir cuatro esquemas distintos, se duda qué esquema seguir: aditivo, recesivo, dominante o codominante. Se consideran los cuatro.

Modelo aditivo

En este modelo el SNP se considera como variable numérica con tres posibles valores: 0, 1 y 2. Se supone que los niveles se agregan de manera conjunta para modelar los datos. Se considera el modelo con `ND23_DIFF` como variable respuesta y `Center`, `Term`, `Gender`, `DBP`, `VLDL_TG`, `Race` y el SNP como predictores.

Se realiza un Q-Q plot de los p-valores asociados al coeficiente de cada polimorfismo que contrastan si este coeficiente es igual o no a 0. Deberían ajustarse a una distribución uniforme si la hipótesis nula se cumple para todos los polimorfismos. Tal y como se ha hecho anteriormente, se grafican con $-log_{10}$ para centrarnos más en la primera parte de los valores (Figura 29).

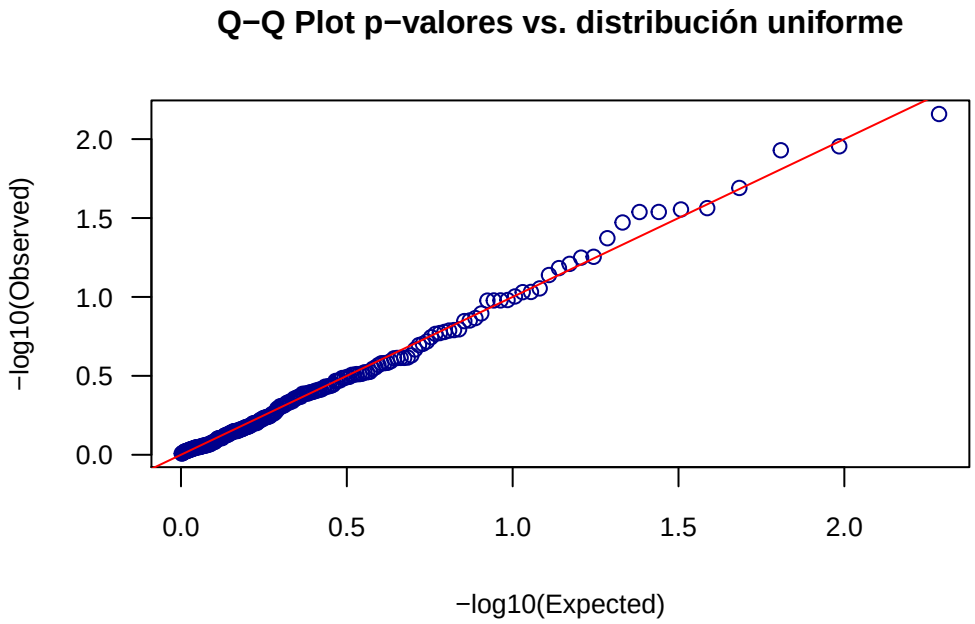

Figura 29: Q-Q Plot de los p-valores de todos los SNP del modelo aditivo para la ganancia del test de fuerza isométrica.

Se puede observar que los p-valores se ajustan bastante bien a una distribución uniforme en casi todos los valores.

A continuación, se realizará un gráfico con todos los p-valores asociados a los coeficientes del parámetro de cada polimorfismo (Figura 30). Se graficarán también dos rectas horizontales indicando el valor de crítico 0.05 y 0.05/193 con la Corrección de Bonferroni, con tal de evitar el falso positivo. Se realiza un segundo gráfico con los p-valores ajustados con el método *False Discovery Rate*.

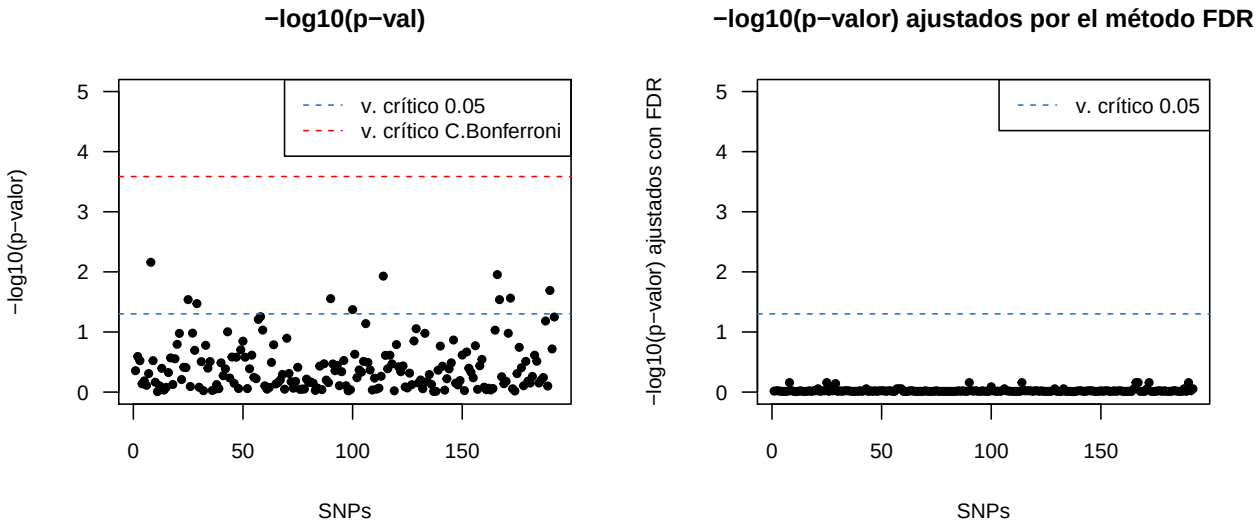

Figura 30: Gráfico de los p-valores de los SNP en el modelo aditivo para la ganancia en el test de fuerza isométrica representando el v. crítico 0.05 y el v. crítico para la C. de Bonferroni y gráfico de los p-valores ajustados por el método FDR.

Puede observarse en el primer gráfico que hay 10 polimorfismos por encima de la recta azul, por lo tanto

SNP	p-valor
adrb2 1042713	0.007
resistin g540a	0.011
kchj11 rs5219	0.012
vdr bsm1	0.02
rs11630261	0.027
gs s287nga	0.028
akt1 c832g c3359g	0.029
resistin c980g	0.029
akt1 g4362c	0.034
igf1 t1245c	0.042

Tabla 2: Lista de los diez polimorfismos con un p-valor más pequeño con el modelo aditivo para el test de ganancia de fuerza isométrica

con un p-valor asociado al coeficiente del SNP inferior a 0.05. Si consideráramos este valor como nivel de significación, podría rechazarse la hipótesis nula que contrasta si el coeficiente es igual a 0. Podría decirse que estos 10 polimorfismos influyen, por separado, en la ganancia de fuerza de bíceps.

Si se tiene en cuenta que al considerar 193 contrastes la posibilidad de falso positivos es alta, puede realizarse la Corrección de Bonferroni. El nivel de significación con dicha corrección es $0.05/193 = 0.0002591$. Ninguno de los p-valores es inferior a esta cifra. Es sabido, que la Corrección de Bonferroni es muy conservadora. También se tiene en cuenta el *False Discovery Rate* para ver la proporción de test significativos que realmente lo son. Utilizando la función de R `FDR` de la versión 1.9 del paquete `astsa` con un máximo de falsos positivos del 5% y la función `p.adjust` para poder graficar los p-valores ajustados por el método tampoco da p-valores significativos.

La lista de los diez polimorfismos con un p-valor más pequeño se encuentra en la Tabla 2.

Seguidamente se considerará el polimorfismo "adrb2_1042713" que es el que tiene el coeficiente asociado más significativo de los anteriores. También se observa que la lista cuenta con algunos polimorfismos que pertenecen al mismo gen ("akt1" y "resistin").

Se representa gráficamente el SNP "adrb2_1042713" respecto la variable respuesta en cada uno de sus niveles (Figura 31). Se tienen 250 genotipos "GG", 269 "GA" y 113 "AA". En el segundo gráfico parace apreciarse que la ganancia de fuerza isométrica en "AA" parece ser ligeramente superior respecto "GA" y "GA" respecto "GG". En estos gráficos, las covariables del modelo no se tienen en cuenta.

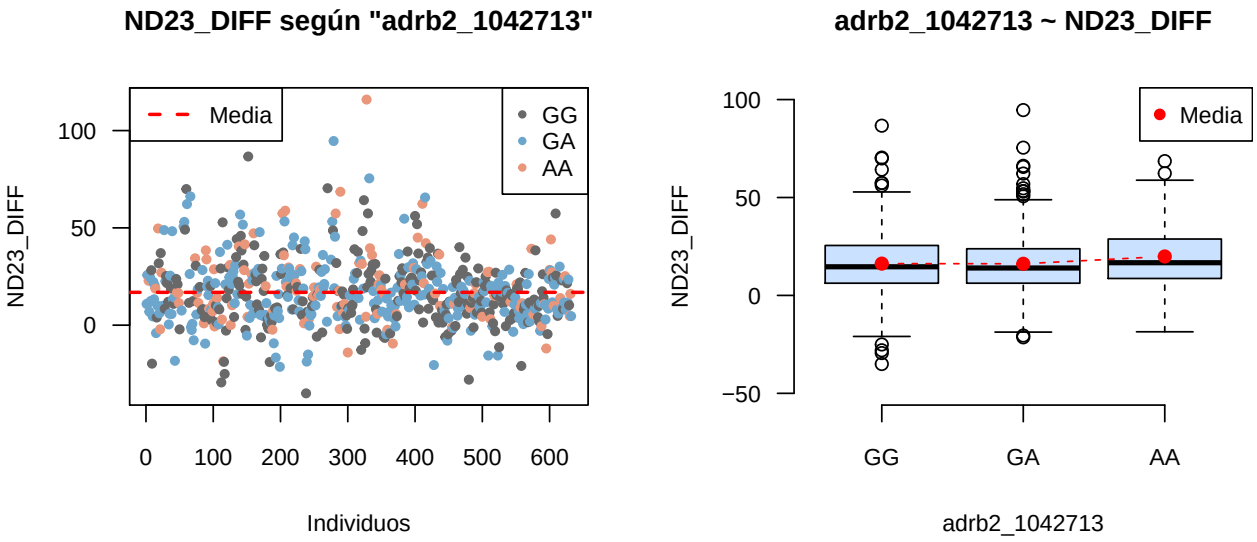

Figura 31: Ganancia de fuerza isométrica según el SNP "adrb2 1042713" con la media con el modelo aditivo.

Se considera el modelo con `ND23_DIFF` como variable respuesta y "adrb2_1042713", `Center`, `Term`, `Gender`, `DBP`, `VLDL_TG` y `Race`, donde con "adrb2_1042713" tiene 3 valores numéricos (0 para GG, 1 para GA y 2 AA)

```
Call:
lm(formula = ND23_DIFF ~ as.numeric(adrb2_1042713) + Center +
    Term + Gender + DBP + VLDL_TG + Race, data = data2)

Residuals:
    Min      1Q  Median      3Q     Max
-51.862  -7.041  -0.725   6.929  77.643

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                10.43983    8.21061   1.272  0.20448
as.numeric(adrb2_1042713) -1.87915    0.69366  -2.709  0.00711 **
CenterFL                  -2.63990    2.63299  -1.003  0.31680
CenterHH                  -3.17592    3.21180  -0.989  0.32350
CenterIR                  10.80524   14.73766   0.733  0.46399
CenterMA                   8.14821    3.08891   2.638  0.00875 **
CenterMI                  -5.54119    3.30410  -1.677  0.09451 .
CenterUC                   8.15232    3.40823   2.392  0.01734 *
CenterWV                   6.98867    2.73460   2.556  0.01106 *
Term02-2                   3.72576    4.48424   0.831  0.40668
Term02-3                   1.95161    2.67226   0.730  0.46573
Term03-1                  -7.56785    2.33460  -3.242  0.00131 **
Term03-2                  -3.87673    4.03160  -0.962  0.33699
Term03-3                  -4.58995    2.46164  -1.865  0.06316 .
GenderMale                 8.61569    1.69091   5.095 5.98e-07 ***
DBP                        0.22539    0.09186   2.454  0.01467 *
VLDL_TG                   -0.16885    0.08061  -2.095  0.03699 *
RaceAsian                 -5.40288    4.08563  -1.322  0.18698
RaceCaucasian             -3.38573    3.25959  -1.039  0.29973
RaceHispanic              -5.94989    4.28415  -1.389  0.16586
```

```
RaceOther                  -1.61056    4.74572  -0.339  0.73455
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.87 on 318 degrees of freedom
  (293 observations deleted due to missingness)
Multiple R-squared:  0.2659,    Adjusted R-squared:  0.2197
F-statistic: 5.759 on 20 and 318 DF,  p-value: 8.391e-13
```

Validación modelo final

Para asegurarnos de que es un buen modelo, procederemos a su validación mediante los residuos, que deberían ser normales y homocedásticos.

En primer lugar, se graficarán los residuos vs los valores ajustados. La nube de puntos debería ser aleatoria y estar centrada en 0. A continuación, se mostrará un Q-Q Plot para ver si se ajustan adecuadamente a la distribución normal (Figura 32).

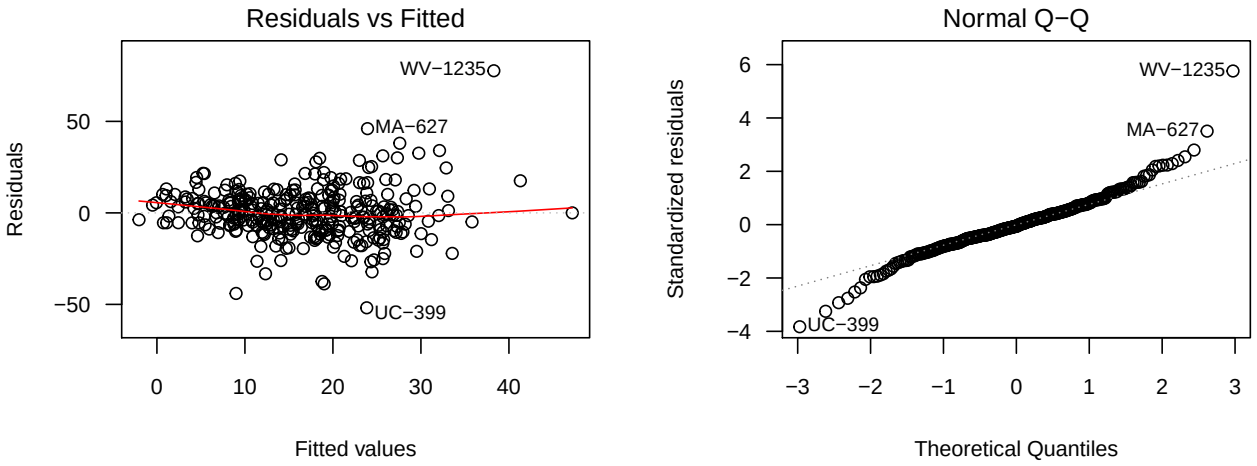

Figura 32: Gráfico residuos vs valores ajustados y QQ-Plot de los residuos del modelo aditivo final para la ganancia de fuerza con las covariables y el SNP "adrb2 1042713" en el modelo aditivo.

En el gráfico de Residuos vs Valores Ajustados se puede observar una nube bastante aleatoria y centrada en 0. No obstante, hay un individuo "WV-1235", que tiene un residuo muy superior al resto. En el Q-Q Plot parece que toda la parte central del grafico se adecúa muy bien a la recta, las colas parecen distar de la misma un poco más pero no es suficiente para afirmar que los residuos no son normales. Lo que sí que llama la atención es que el individuo "WV-1235" se aleja mucho de la normalidad. Puesto que el sujeto no cumple ninguna de las dos hipótesis, se reestimará el modelo definitivo sin el mismo.

```
Call:
lm(formula = ND23_DIFF ~ as.numeric(adrb2_1042713) + Center +
    Term + Gender + DBP + VLDL_TG + Race, data = data2[-328,
    ])

Residuals:
    Min      1Q  Median      3Q     Max
-51.293  -7.080  -0.139   7.181  44.977

Coefficients:
```

```
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               13.36430    7.79741   1.714  0.08752 .
as.numeric(adrb2_1042713) -1.48470    0.66069  -2.247  0.02532 *
CenterFL                  -3.16544    2.49721  -1.268  0.20588
CenterHH                  -3.32669    3.04446  -1.093  0.27535
CenterIR                  11.92267   13.97052   0.853  0.39407
CenterMA                   7.72085    2.92872   2.636  0.00880 **
CenterMI                  -5.91505    3.13244  -1.888  0.05990 .
CenterUC                   7.26017    3.23387   2.245  0.02545 *
CenterWV                   5.34337    2.60613   2.050  0.04116 *
Term02-2                   4.47441    4.25224   1.052  0.29349
Term02-3                   0.26170    2.54816   0.103  0.91827
Term03-1                  -7.25844    2.21347  -3.279  0.00116 **
Term03-2                  -3.69773    3.82153  -0.968  0.33398
Term03-3                  -4.71662    2.33340  -2.021  0.04408 *
GenderMale                 7.90657    1.60700   4.920 1.39e-06 ***
DBP                        0.17643    0.08744   2.018  0.04446 *
VLDL_TG                   -0.14591    0.07650  -1.907  0.05739 .
RaceAsian                 -4.87563    3.87360  -1.259  0.20907
RaceCaucasian             -3.61004    3.08987  -1.168  0.24354
RaceHispanic              -5.82055    4.06085  -1.433  0.15275
RaceOther                 -1.67882    4.49832  -0.373  0.70924
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.15 on 317 degrees of freedom
  (293 observations deleted due to missingness)
Multiple R-squared:  0.2539,    Adjusted R-squared:  0.2068
F-statistic: 5.394 on 20 and 317 DF,  p-value: 8.276e-12
```

El R^2_{adj} del modelo es de 0.2068, cosa que indica que el modelo explica un 20.68% de la variabilidad de la ganancia de fuerza de los individuos. El p-valor asociado al estadístico F que contrasta la utilidad del modelo es de 8.276e-12, más pequeño que el valor crítico 0.05. Se puede rechazar la hipótesis nula y decir que existe relación entre la variable respuesta y los predictores, es un modelo útil. Tanto el sexo como el SNP en cuestión son las dos variables más influyentes. Parece los individuos que tienden a ganar más fuerza isométrica son los hombres, las personas examinadas en *University of West Virginia*, en *University of Massachusetts* y *Central Michigan University* y las personas con presión sanguínea diastólica alta. Contariamente, en media, las personas con muchas lipoproteínas de muy baja densidad altas y examinadas en pierden fuerza después del entrenamiento. Los individuos con el genotipo "GG" en el gen "adrb2_1042713" son las que, en media, más fuerza isométrica han ganado seguidas de las que tienen el genotipo "GA". Cuando se aumenta "A" en una unidad, la ganancia de fuerza isométrica disminuye, en media, en 1.49 unidades. Esto parece ser contario a lo que se observaba en el gráfico dela Figura 31, sin embargo, el resultado del modelo es más fiable ya que se tienen en cuenta también las covariables inlfuyentes.

Modelo recesivo

En este modelo el SNP se considera como variable numérica con dos posibles valores: 0 para el homocigoto más común y el heterocigoto y 1 para el homocigoto menos común. Se considera el modelo con recesivo `ND23_DIFF` como variable respuesta y `Center`, `Term`, `Gender`, `DBP`, `VLDL_TG`, `Race` y el SNP, con los dos valores explicados anteriormente.

Mediante un Q-Q plot de los p-valores asociados al coeficiente de cada polimorfismo que contrastan si este coeficiente es igual a 0 se verá si se ajustan a una distribución uniforme si la hipótesis nula se cumple para todos los polimorfismos. Como se ha hecho anteriormente, se grafican con $-log_{10}$ para centrarnos más en la primera parte de los valores (Figura 33).

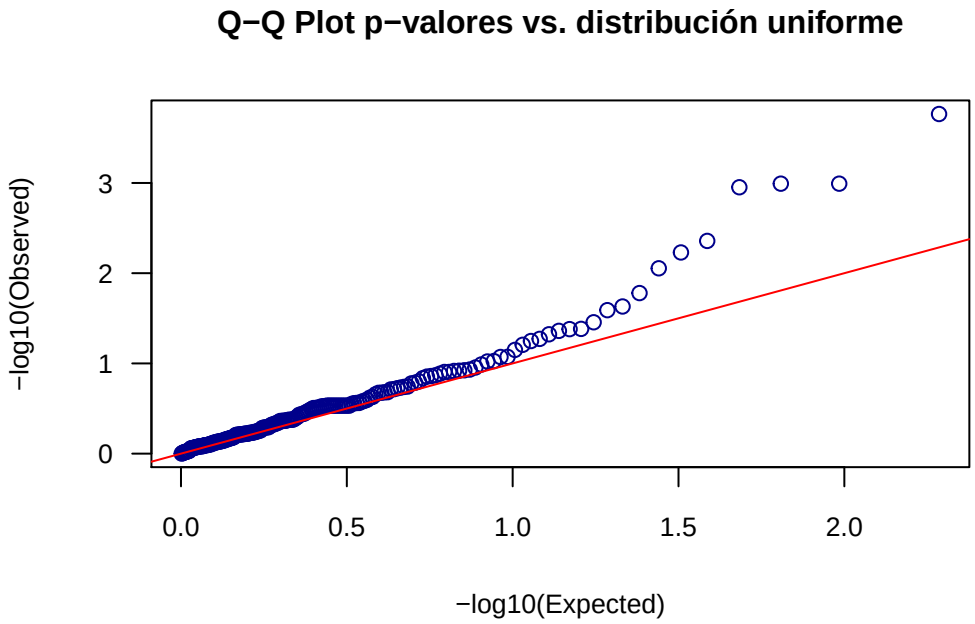

Figura 33: Q-Q Plot de los p-valores de todos los SNP del modelo recesivo para la ganancia del test de fuerza isométrica.

La cola de la izquierda se ajusta bastante bien a una distribución uniforme como puede observarse. En cambio, en la cola de la derecha parece haber una serie de valores que distan bastante de esta distribución, es probable que haya p-valores muy significativos.

Seguidamente, se realizará un gráfico con todos los p-valores asociados a los coeficientes del parámetro de cada SNP (Figura 34). Se mostrarán también dos rectas horizontales que indicarán el valor de crítico 0.05 y 0.05/193 con la Corrección de Bonferroni, con tal de evitar el falso positivo y un segundo gráfico con los p-valores ajustados con el método FDR.

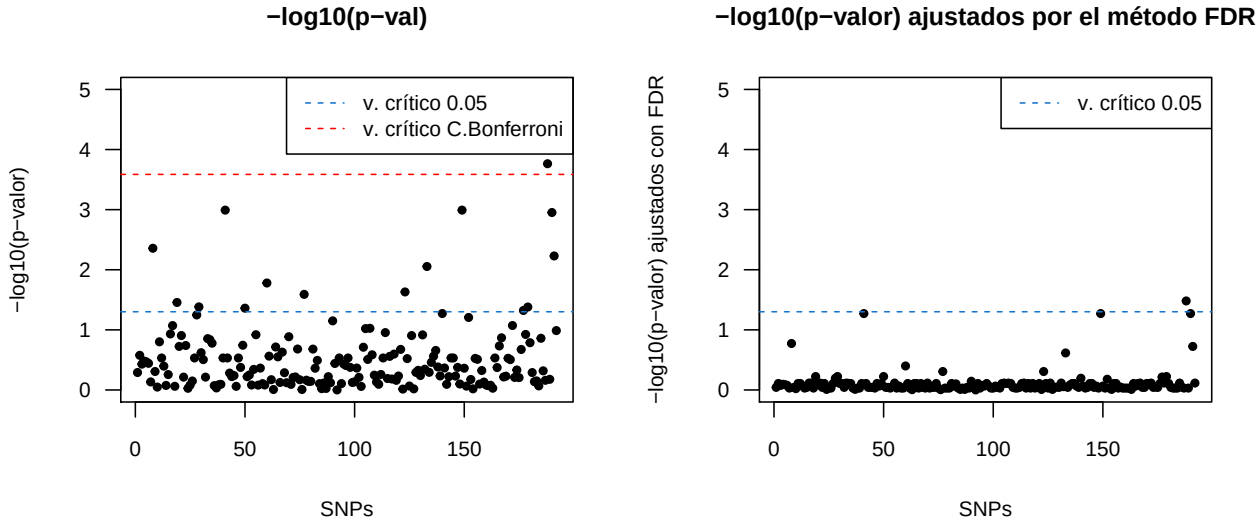

Figura 34: Gráfico de los p-valores de los SNP en el modelo recesivo para la ganancia en el test de fuerza isométrica representando el v. crítico 0.05 y el v. crítico para la C. de Bonferroni y gráfico de los p-valores ajustados por el método FDR.

SNP	p-valor
vdr taq1	0
ankrd6 q122e	0.001
pcr5 snp1	0.001
vdr bsm1	0.001
adrb2 1042713	0.004
vdr rs731236	0.006
nr3c1 rs10482614	0.009
cast rs7724759	0.017
mgst3 4147542	0.023
fbox32 rs3739287	0.026

Tabla 3: Lista de los diez polimorfismos con un p-valor más pequeño con el modelo recesivo para el test de ganancia de fuerza isométrica

Se ha encontrado un SNP muy significativo, que está por encima de la línea roja que marca el valor crítico con la Corrección de Bonferroni. Este polimorfismo parece tener mucha importancia en la ganancia de fuerza isométrica. Además, se tienen 13 SNPs más por encima del valor crítico 0.05, que también parecen tener relación con la variable respuesta aunque también debe tenerse en cuenta que la probabilidad de falso positivo cuando se realizan 193 test es alta.

También se tiene en cuenta el *False Discovery Rate* como en los apartados anteriores. Se utiliza la función de R `FDR` de la versión 1.9 del paquete **`astsa`** con un máximo de falsos positivos del 5% y se grafican los p-valores obtenidos de ajustarlos con la función de R `p.adjust` de la versión 3.6.0 del paquete **`stats`**. El único p-valor significativo el número 188, el mismo que la Corrección de Bonferroni. No obstante, en el gráfico puede observarse que hay tres valores en el límite.

La lista de los diez polimorfismos con un p-valor más pequeño se encuentra en la Tabla 3.

Se estudiará más a fondo el SNP más significativo: “vdr_taq1”. Otro aspecto a destacar, es que hay dos polimorfismos significativos del gen “vdr” y “akt1”.

Se representa gráficamente el polimorfismo “vdr_taq1” respecto a la variable respuesta en cada uno de sus niveles (Figura 35). En estos gráficos, parece que “TT” y “TC” tengan una medida un poco superior a “CC”. Este SNP tiene 66 “CC” genotipos, 266 “CT” y 242 “TT”.

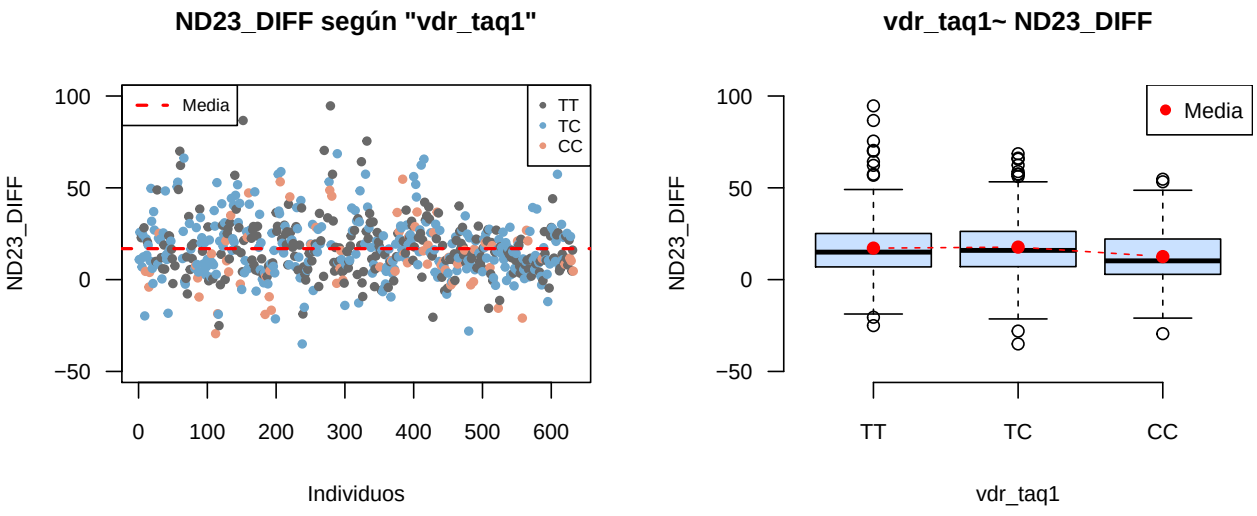

Figura 35: Test de ganancia de fuerza isométrica según el SNP "vdr taq1" con la media.

Se considera el modelo con ND23_DIFF como variable respuesta y "vdr_taq1", Center, Term, Gender, DBP, VLDL_TG y Race, donde con "'vdr_taq1" es una variable numérica con dos niveles. El nivel basal incluirá los niveles del factor "TT" y "TC", y la variable el nivel "CC".

```
Call:
lm(formula = ND23_DIFF ~ as.numeric(vdr_taq1b) + Center + Term +
    Gender + DBP + VLDL_TG + Race, data = data2)

Residuals:
    Min      1Q  Median      3Q     Max
-50.369  -8.399  -0.735   7.360  79.202

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            5.11869    7.70814   0.664 0.506999
as.numeric(vdr_taq1b) -8.64668    2.28183  -3.789 0.000172 ***
CenterFL              -2.24995    2.75199  -0.818 0.414047
CenterHH              -0.22632    3.32972  -0.068 0.945841
CenterIR              29.21766   15.61861   1.871 0.062054 .
CenterMA               8.23772    3.22269   2.556 0.010920 *
CenterMI              -5.49045    2.66796  -2.058 0.040189 *
CenterUC               3.62934    2.71193   1.338 0.181497
CenterWV               9.42374    2.87113   3.282 0.001112 **
Term02-2              -0.20220    2.94630  -0.069 0.945317
Term02-3               3.36074    2.26897   1.481 0.139279
Term03-1              -4.84768    2.04734  -2.368 0.018328 *
Term03-2              -3.84659    3.99590  -0.963 0.336262
Term03-3              -2.19213    2.17062  -1.010 0.313095
GenderMale             7.64931    1.59017   4.810 2.08e-06 ***
DBP                    0.18819    0.08466   2.223 0.026743 *
VLDL_TG               -0.15356    0.07485  -2.052 0.040790 *
RaceAsian             -5.47302    4.00549  -1.366 0.172520
RaceCaucasian         -1.56084    3.25155  -0.480 0.631446
RaceHispanic          -7.69466    4.38504  -1.755 0.080001 .
RaceOther              2.78914    4.83824   0.576 0.564588
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.05 on 438 degrees of freedom
  (173 observations deleted due to missingness)
Multiple R-squared:  0.2161,    Adjusted R-squared:  0.1803
F-statistic: 6.036 on 20 and 438 DF,  p-value: 3.299e-14
```

Validación modelo final

Se quiere asegurar que sea un buen modelo, por lo que se procederá a su validación mediante los residuos, que deberían ser normales y homocedásticos.

Primeramente, se graficarán los residuos vs los valores ajustados. La nube de puntos debería ser aleatoria y estar centrada en 0. Para continuar, se mostrará un Q-Q Plot para ver si se ajustan adecuadamente a la distribución normal (Figura 36).

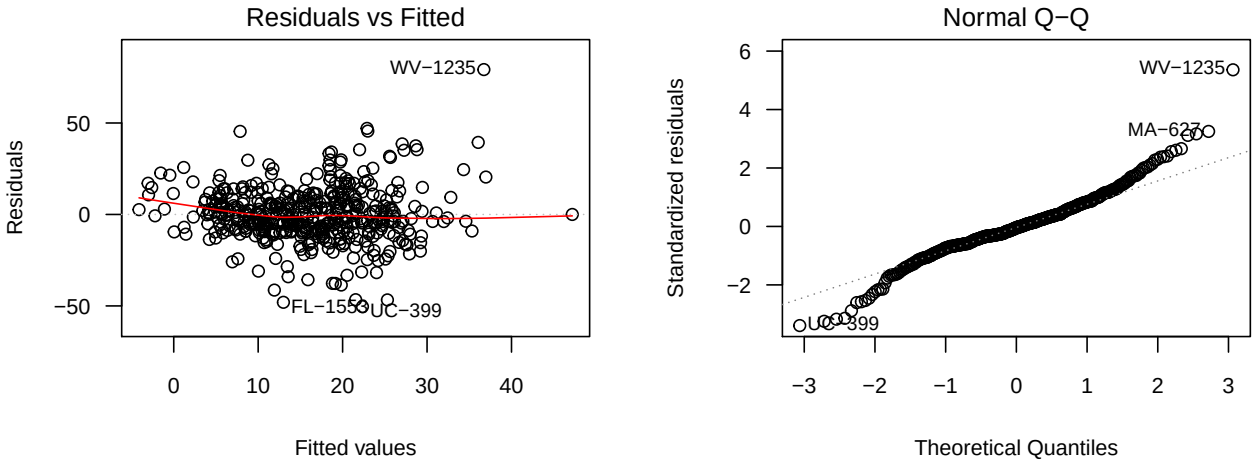

Figura 36: Gráfico residuos vs valores ajustados y QQ-Plot de los residuos del modelo recesivo final para la ganancia de fuerza con las covariables y el SNP "vdr taq1" en el modelo recesivo.

En el gráfico de Residuos vs Valores Ajustados se puede ver una nube que parece aleatoria y centrada en 0. Sin embargo, hay el punto del individuos "WV-1235", que tiene un residuo muy superior al resto. En el Q-Q Plot parece que toda la parte del centro del gráfico se adecúa muy bien a la recta aunque las colas parecen distar de la misma un poco más. No obstante, no es suficiente para afirmar que los residuos no son normales. Lo que sí que llama más la atención es que el individuo "WV-1235" se aleja mucho de la normalidad. Como en el modelo anterior, a causa de que el individuo no cumple ninguna de las dos hipótesis, se reestimará el modelo definitivo sin el mismo.

```
Call:
lm(formula = ND23_DIFF ~ as.numeric(vdr_taq1b) + Center + Term +
    Gender + DBP + VLDL_TG + Race, data = data2[-328, ])

Residuals:
    Min      1Q  Median      3Q     Max
-49.906  -8.047  -0.715   7.048  46.812

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           7.96105    7.47726   1.065 0.287599
as.numeric(vdr_taq1b) -8.33524    2.20898  -3.773 0.000183 ***
CenterFL              -2.62486    2.66413  -0.985 0.325042
CenterHH              -0.47878    3.22269  -0.149 0.881964
CenterIR              29.62097   15.11523   1.960 0.050669 .
CenterMA               7.99407    3.11909   2.563 0.010713 *
CenterMI              -5.65244    2.58211  -2.189 0.029119 *
CenterUC               3.41491    2.62478   1.301 0.193935
CenterWV               7.91092    2.79196   2.833 0.004818 **
Term02-2              -0.09357    2.85137  -0.033 0.973836
Term02-3               2.35932    2.20325   1.071 0.284834
Term03-1              -4.74048    1.98142  -2.392 0.017157 *
Term03-2              -3.52944    3.86749  -0.913 0.361960
Term03-3              -2.26635    2.10068  -1.079 0.281243
GenderMale             7.11663    1.54190   4.615 5.16e-06 ***
DBP                    0.15432    0.08216   1.878 0.061023 .
VLDL_TG               -0.13926    0.07248  -1.921 0.055334 .
```

```
RaceAsian             -5.23711    3.87658  -1.351 0.177407
RaceCaucasian         -1.73505    3.14687  -0.551 0.581671
RaceHispanic          -7.61115    4.24369  -1.794 0.073581 .
RaceOther              2.80036    4.68225   0.598 0.550097
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.56 on 437 degrees of freedom
  (173 observations deleted due to missingness)
Multiple R-squared:  0.2044,    Adjusted R-squared:  0.168
F-statistic: 5.614 on 20 and 437 DF,  p-value: 5.635e-13
```

El R^2_{adj} del modelo es de 0.168, indica que el modelo explica un 16.8% de la variabilidad de la ganancia de fuerza isométrica de los individuos. El p-valor asociado al estadístico F que contrasta la utilidad del modelo es de 5.635e-13, inferior al valor crítico 0.05. Se puede rechazar la hipótesis nula y decir que existe relación entre la variable respuesta y las variables explicativas, es un modelo útil. Parece los individuos que tienden a ganar más fuerza son los hombres, las personas examinadas en *University of West Virginia* y en *University of Massachusetts*. En *Central Michigan University* y en el trimestre "03-1" parece que se tiende a ganar menos fuerza que en el resto de universidades. Los individuos con el genotipo "TT" y "TC" en el gen "vdr_taq1" son las que, en media, más fuerza isométrica ganan, concuerda con los resultados gráficos. Si se tiene el genotipo "CC", la ganancia de fuerza isométrica disminuye, en media, en 8.33 unidades.

Modelo dominante

En este modelo el SNP se considera como variable numérica con dos posibles valores: 0 para el homocigoto más común y 1 para heterocigoto y el homocigoto menos común. Se considera el modelo con recesivo `ND23_DIFF` como variable respuesta y `Center`, `Term`, `Gender`, `DBP`, `VLDL_TG`, `Race` y el SNP, con los dos valores explicados anteriormente.

Se vuelve a realizar un Q-Q plot de los p-valores asociados al coeficiente de cada polimorfismo que contrastan si este coeficiente es igual o no a 0. Deberían ajustarse a una distribución uniforme si la hipótesis nula es cierta para todos los valores (Figura 37).

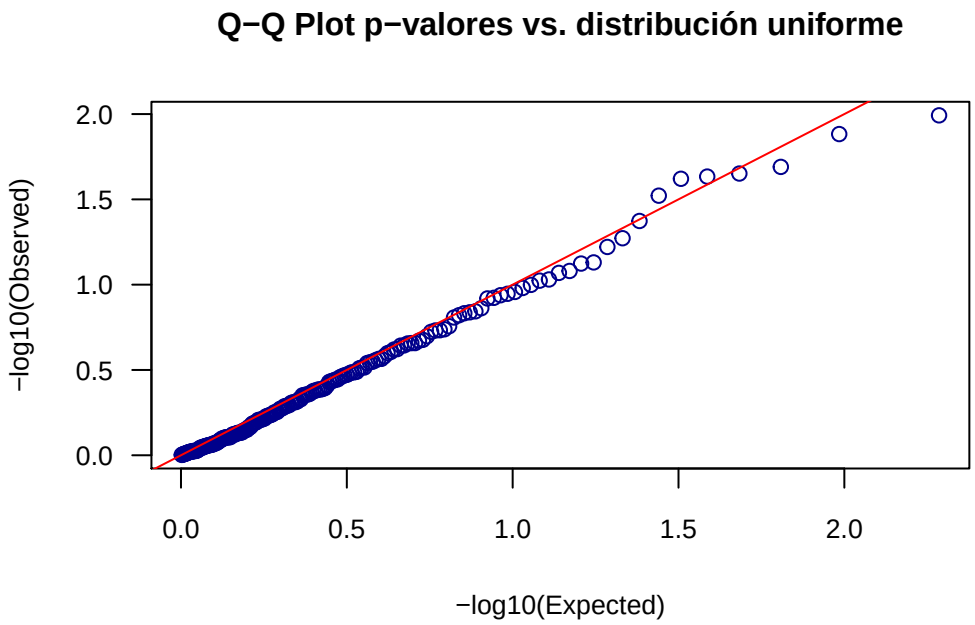

Figura 37: Q-Q Plot de los p-valores de todos los SNP del modelo dominante para la ganancia del test de fuerza isométrica.

SNP	p-valor
resistin g540a	0.01
igf1 t1245c	0.013
kchj11 rs5219	0.02
carp c105t	0.022
akt1 c832g c3359g	0.023
carp a8470g	0.024
cast rs754615	0.03
resistin c980g	0.042
rs11630261	0.053
gs s287nga	0.06

Tabla 4: Lista de los diez polimorfismos con un p-valor más pequeño con el modelo dominante para el test de ganancia de fuerza isométrica

No parece que los valores disten excesivamente de la recta roja que marca la uniformidad, no se puede decir que los p-valores no sigan una distribución uniforme.

A continuación, se realizará un gráfico con los p-valores asociados a los coeficientes del parámetro de cada SNP (Figura 38). Además, se mostrarán dos rectas horizontales que indicarán el valor de crítico 0.05 y 0.05/193 con la Corrección de Bonferroni, con tal de evitar el error de tipo I. En el segundo gráfico se muestran los p-valores ajustados según el método FDR.

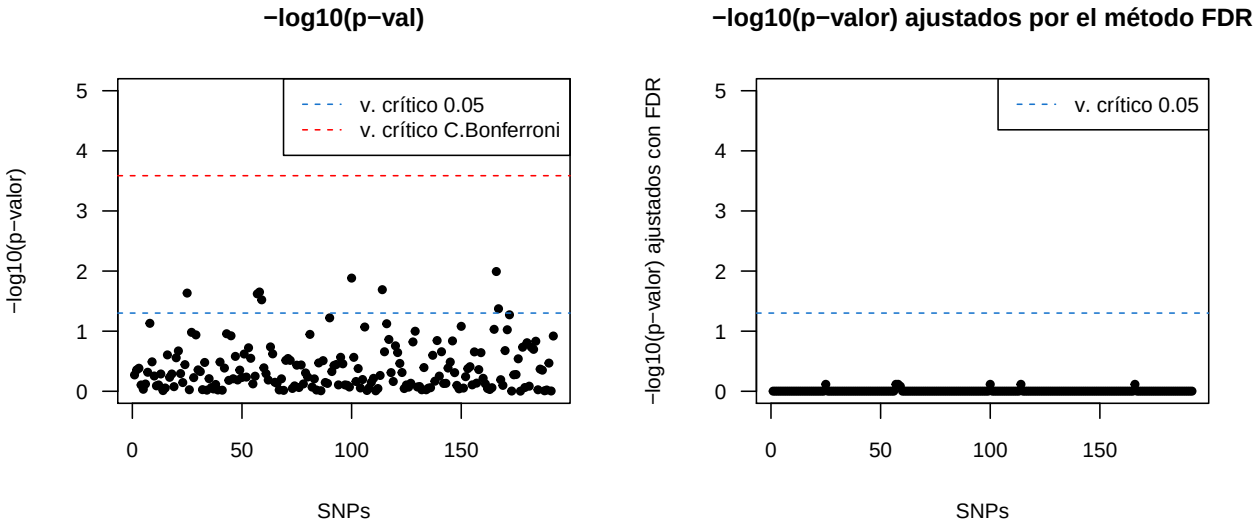

Figura 38: Gráfico de los p-valores de los SNP en el modelo dominante para la ganancia en el test de fuerza isométrica representando el v. crítico 0.05 y el v. crítico para la C. de Bonferroni y gráfico de los p-valores ajustados por el método FDR.

No hay ningún valor por encima de la recta roja que marca el valor crítico de $0.05/193 = 0.0002591$ obtenido mediante la Corrección de Bonferroni. Sin embargo, sí hay 8 p-valores por encima del valor crítico 0.05. El *False Discovery Rate* aunque no es tan conservador como el método de Bonferroni tampoco da ningún SNP significativo.

La lista de los diez polimorfismos con un p-valor más pequeño se puede encontrar en la Tabla 4.

Se estudiará más a fondo el SNP más significativo: "resistin_g540a". Otro aspecto que se puede destacar, es que hay dos polimorfismos significativos del gen "resistin".

Se representa gráficamente el polimorfismo “resistin_g540a” respecto a la variable respuesta en cada uno de sus niveles (Figura 39). Gráficamente no parece que haya grandes diferencias entre grupos. Este SNP tiene 63 “AA” genotipos, 244 “GA” y 307 “GG”.

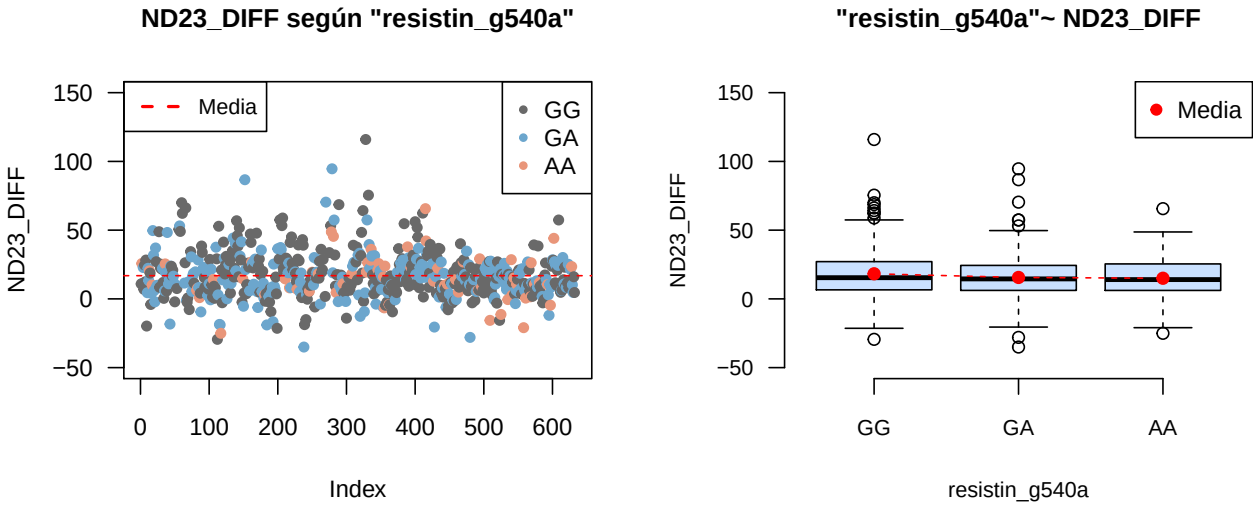

Figura 39: Test de ganancia de fuerza isométrica según el SNP "resistin a537c" con la media.

Se considera el modelo con `ND23_DIFF` como variable respuesta y “resistin_g540a”, `Center`, `Term`, `Gender`, `DBP`, `VLDL_TG` y `Race`, donde con resistin_g540a es una variable numérica con dos niveles. El nivel basal incluirá los niveles del factor “GG” y la variable el nivel “GA” y “AA”.

```
Call:
lm(formula = ND23_DIFF ~ as.numeric(resistin_g540ac) + Center +
    Term + Gender + DBP + VLDL_TG + Race, data = data2)

Residuals:
    Min      1Q  Median      3Q     Max
-51.067  -7.560  -1.093   8.521  78.803

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  9.41690    7.85760   1.198  0.23139
as.numeric(resistin_g540ac) -3.81300    1.47732  -2.581  0.01017 *
CenterFL                    -2.92699    2.76979  -1.057  0.29121
CenterHH                    -1.19115    3.36518  -0.354  0.72354
CenterIR                    20.01052   15.65250   1.278  0.20178
CenterMA                     7.72769    3.24730   2.380  0.01775 *
CenterMI                    -6.40123    2.67993  -2.389  0.01734 *
CenterUC                     3.36534    2.73514   1.230  0.21920
CenterWV                     8.20652    2.88406   2.845  0.00464 **
Term02-2                    -1.02091    2.96973  -0.344  0.73118
Term02-3                     3.60813    2.28696   1.578  0.11536
Term03-1                    -5.14500    2.06455  -2.492  0.01307 *
Term03-2                    -3.04422    4.02694  -0.756  0.45008
Term03-3                    -1.83146    2.19101  -0.836  0.40367
GenderMale                   7.77499    1.60713   4.838 1.82e-06 ***
DBP                          0.18158    0.08536   2.127  0.03397 *
```

```
VLDL_TG                      -0.16065    0.07548  -2.128  0.03386 *
RaceAsian                    -6.86380    4.07943  -1.683  0.09318 .
RaceCaucasian                -3.91937    3.36214  -1.166  0.24436
RaceHispanic                -10.22703    4.49856  -2.273  0.02349 *
RaceOther                     0.82945    4.92728   0.168  0.86640
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.18 on 438 degrees of freedom
  (173 observations deleted due to missingness)
Multiple R-squared:  0.2025,    Adjusted R-squared:  0.1661
F-statistic: 5.561 on 20 and 438 DF,  p-value: 8.006e-13
```

Validación del modelo final

Para asegurarnos de que es un buen modelo, procederemos a su validación mediante los residuos, que deberían ser normales y homocedásticos.

En primer lugar, se graficarán los residuos vs los valores ajustados. Como se ha dicho anteriormente, la nube de puntos debería ser aleatoria y estar centrada en 0. Seguidamente, se mostrará un Q-Q Plot para ver si se ajustan adecuadamente a la distribución normal (Figura 40).

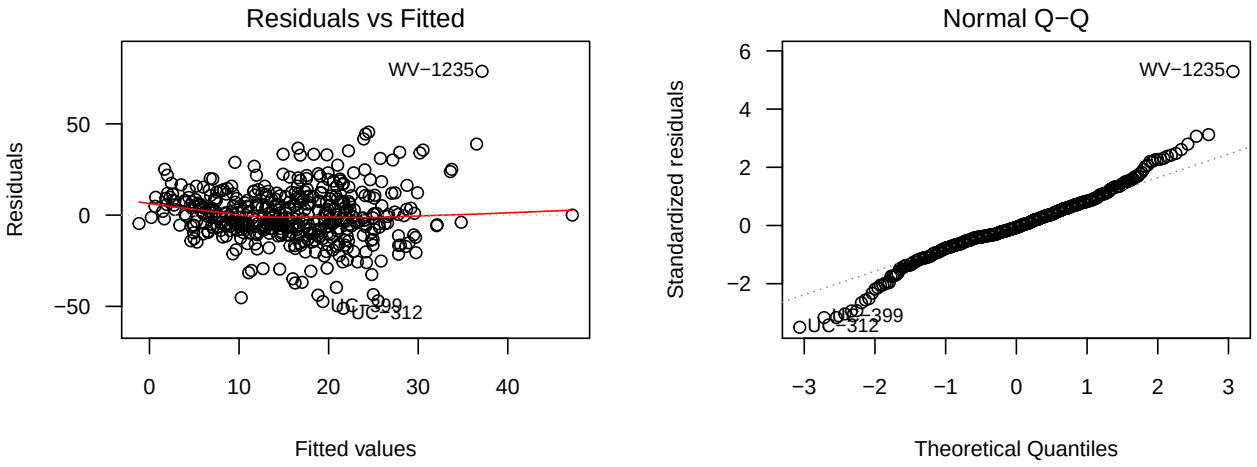

Figura 40: Gráfico residuos vs valores ajustados y QQ-Plot de los residuos del modelo dominante final para el test de ganancia de fuerza con las covariables y el SNP "resistin g540a" en el modelo.

En el gráfico de Residuos vs Valores Ajustados se puede observar una nube bastante aleatoria y centrada en 0. No obstante igual que en el modelo anterior, el individuo “WV-1235”, que tiene un residuo muy superior al resto. En el Q-Q Plot parece que toda la parte central del grafico se adecúa muy bien a la recta, las colas parecen distar de la misma un poco más, sin embargo, se cree que no es suficiente para afirmar que los residuos no son normales. El individuo “WV-1235” se aleja mucho de la normalidad. Puesto que el sujeto no cumple ninguna de las dos hipótesis, se reestimará el modelo definitivo sin el mismo.

```
Call:
lm(formula = ND23_DIFF ~ as.factor(resistin_g540ac) + Center +
    Term + Gender + DBP + VLDL_TG + Race, data = data2[-328,
    ])

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-50.819  -7.607  -0.752   8.184  45.520

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 11.90724    7.62466   1.562   0.1191
as.factor(resistin_g540ac)1 -3.44198    1.43257  -2.403   0.0167 *
CenterFL                    -3.27655    2.68363  -1.221   0.2228
CenterHH                    -1.37105    3.25974  -0.421   0.6743
CenterIR                    20.86257   15.16210   1.376   0.1695
CenterMA                     7.50136    3.14567   2.385   0.0175 *
CenterMI                    -6.53018    2.59593  -2.516   0.0122 *
CenterUC                     3.16666    2.64955   1.195   0.2327
CenterWV                     6.75456    2.80617   2.407   0.0165 *
Term02-2                    -0.86717    2.87667  -0.301   0.7632
Term02-3                     2.60785    2.22274   1.173   0.2413
Term03-1                    -5.02194    1.99988  -2.511   0.0124 *
Term03-2                    -2.76593    3.90090  -0.709   0.4787
Term03-3                    -1.93293    2.12233  -0.911   0.3629
GenderMale                   7.22206    1.55999   4.630 4.84e-06 ***
DBP                          0.14752    0.08292   1.779   0.0759 .
VLDL_TG                     -0.14606    0.07316  -1.996   0.0465 *
RaceAsian                   -6.48906    3.95200  -1.642   0.1013
RaceCaucasian               -3.88959    3.25663  -1.194   0.2330
RaceHispanic                -9.92159    4.35774  -2.277   0.0233 *
RaceOther                    1.01927    4.77278   0.214   0.8310
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.7 on 437 degrees of freedom
  (173 observations deleted due to missingness)
Multiple R-squared:  0.1892,    Adjusted R-squared:  0.1521
F-statistic: 5.099 on 20 and 437 DF,  p-value: 1.794e-11
```

El R^2_{adj} del modelo es de 0.1521, cosa que indica que el modelo explica un 15.21% de la variabilidad de la ganancia de fuerza de los individuos. El p-valor asociado al estadístico F que contrasta la utilidad del modelo es de 1.794e-11, más pequeño que el valor crítico 0.05. Se puede rechazar la hipótesis nula y decir que existe relación entre la variable respuesta y los predictores, es un modelo útil. Tanto el sexo como el polimorfismo son las dos variables más influyentes. Al parecer, los sujetos que tienden a ganar más fuerza isométrica son los hombres, las personas examinadas en *University of West Virginia* y en *University of Massachusetts* y las personas con el genotipo "GG" en el gen "resistin_g540a". En cambio las personas con muchas lipoproteínas de muy baja densidad altas, las de raza hispánica, las examinadas en *Central Michigan University* y los individuos estudiados en el trimestre "03-1" tienden a perder fuerza después del entrenamiento. La presencia del alelo "A" hace que la ganancia de fuerza isométrica sea, en media, 3.44 unidades menor.

Modelo codominante

En este modelo se considerará la variable respuesta como una variable categórica, en la que se tendrán dos variables indicadoras: una para los heterocigotos y otra para el homocigoto menos común. El nivel basal serán los homocigotos más frecuentes.

Se vuelve a realizar un Q-Q plot de los p-valores asociados al coeficiente de cada SNP que contrastan si este coeficiente es igual a 0. Deberían ajustarse a una distribución uniforme si se cumple la hipótesis nula (Figura 41).

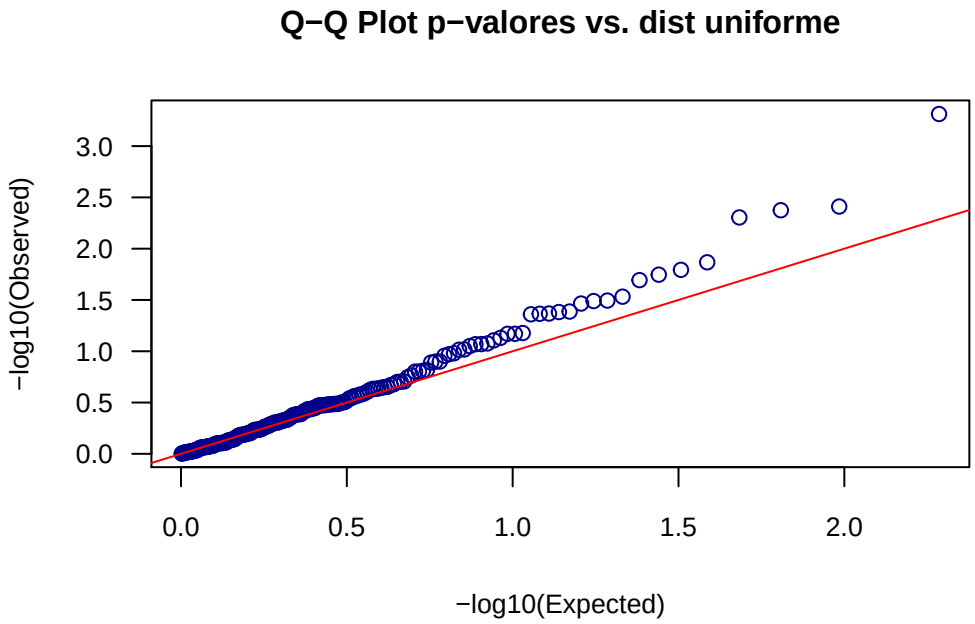

Figura 41: Q-Q Plot de los p-valores de todos los SNP para el esquema codominante utilizando la ganancia de fuerza como variable respuesta.

Se puede observar que los p-valores solamente se ajustan bastante bien a una distribución uniforme al principio en los dos gráficos. Sin embargo, en la parte final se ve que distan bastante de esta distribución. Es probable que haya p-valores significativamente pequeños.

A continuación, se realizará un gráfico con todos los p-valores asociados a los coeficientes del parámetro de cada polimorfismo para cada nivel del factor. De la misma manera que en el apartado anterior, se graficarán también dos rectas horizontales indicando el valor de crítico 0.05 y 0.05/193 con la Corrección de Bonferroni, con tal de evitar el falso positivo y los p-valores ajustados con el *False Discovery Rate* (Figura 42).

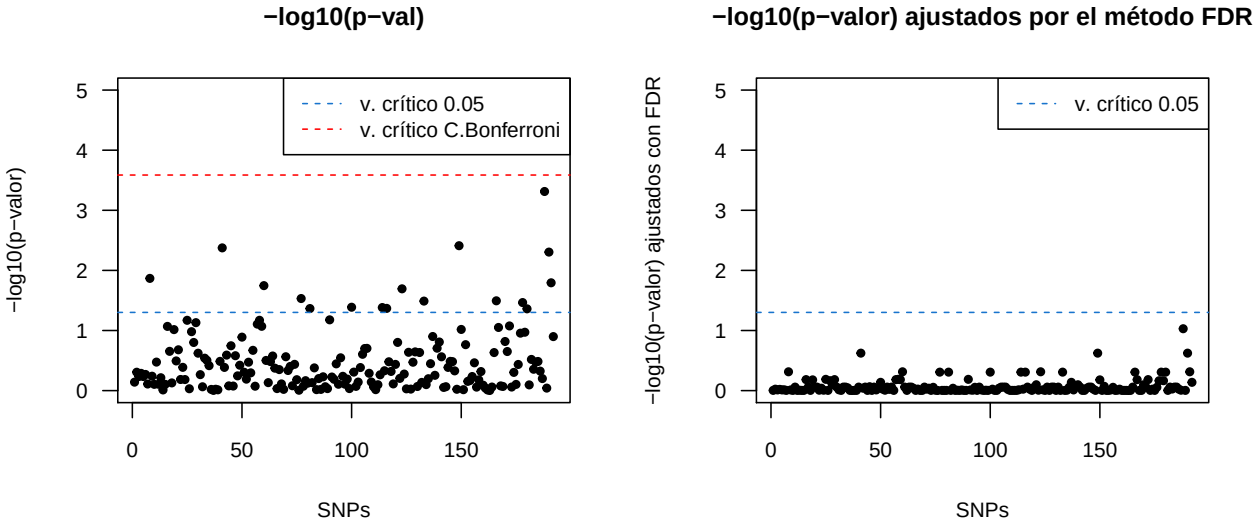

Figura 42: Gráfico de los p-valores de los SNP en el modelo codominante para la ganancia en el test de fuerza isométrica representando el v. crítico 0.05 y el v. crítico para la C. de Bonferroni y gráfico de los p-valores ajustados por el método FDR.

SNP	p-valor
vdr taq1	0
pcr5 snp1	0.004
ankrd6 q122e	0.004
vdr bsm1	0.005
adrb2 1042713	0.014
vdr rs731236	0.016
cast rs7724759	0.018
mgst3 4147542	0.02
fbox32 rs3739287	0.029
resistin g540a	0.032

Tabla 5: Lista de los diez polimorfismos con un p-valor más pequeño para el esquema codominante para el test de ganancia de fuerza isométrica.

En este gráfico puede observarse que hay 17 polimorfismos por encima de la recta azul, por lo tanto con un p-valor asociado al coeficiente del SNP inferior a 0.05. Si consideráramos este valor como nivel de significación, podría rechazarse la hipótesis nula que contrasta si el coeficiente es igual a 0. Podría decirse que estos 17 polimorfismos influyen, por separado, en la ganancia de fuerza de bíceps. También puede observar que hay un SNP por más cerca de la recta roja, valor crítico que establece la Corrección de Bonferroni, es muy significativo, interesará estudiarlo más a fondo. Aunque el *False Discovery Rate* no da ningún p-valor significativo, se puede observar que hay un valor cerca de la línea que marca el valor crítico, el mismo que en el test de Bonferroni.

La lista de los diez polimorfismos con un p-valor más pequeño se encuentra en la Tabla 5.

Seguidamente se considerará el SNP más significativo de la lista: "vdr_taq1". Se observa que hay tres polimorfismos del gen "vdr".

Se representa gráficamente el polimorfismo "vdr_taq1" respecto a la variable respuesta en cada uno de sus niveles (Figura 43). Según estos gráficos, parece que las personas con el genotipo "CC" tienden a ganar menos fuerza isomética. Este SNP tiene 66 "CC" genotipos, 266 "CT" y 242 "TT".

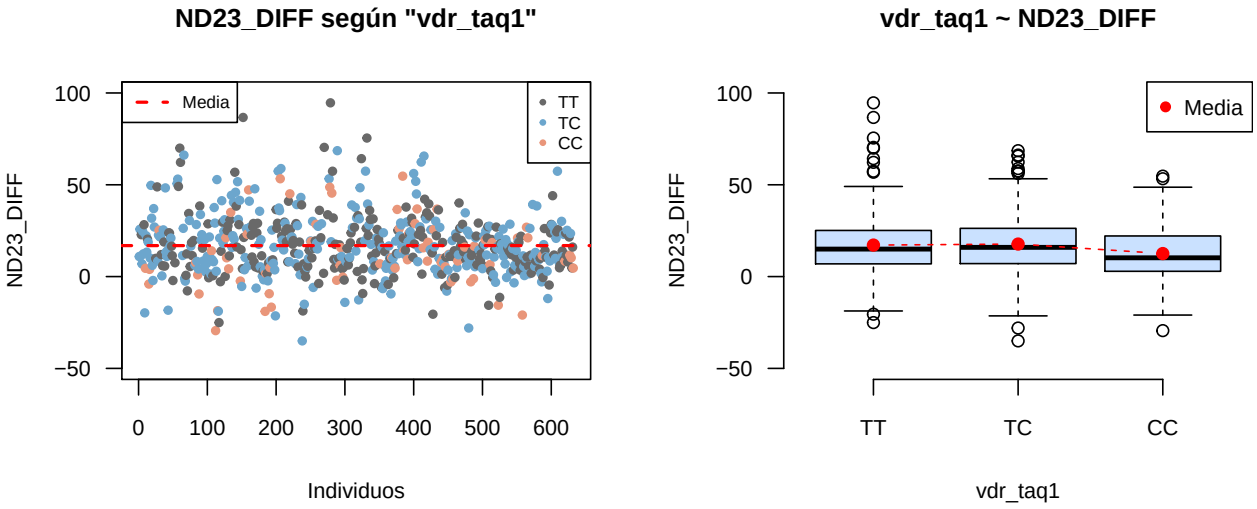

Figura 43: Test de ganancia de fuerza isométrica según el SNP "vdr taq1" con la media.

```
Call:
lm(formula = ND23_DIFF ~ as.factor(vdr_taq1) + Center + Term +
```

```
    Gender + DBP + VLDL_TG + Race, data = data2)

Residuals:
    Min      1Q  Median      3Q     Max
-51.770  -8.140  -0.925   7.887  77.193

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -5.10757    8.35102  -0.612  0.54113
as.factor(vdr_taq1)CT 10.05510    2.43799   4.124 4.49e-05 ***
as.factor(vdr_taq1)TT  8.13494    2.50162   3.252  0.00124 **
CenterFL             -2.49758    2.81503  -0.887  0.37546
CenterHH             -0.42477    3.42767  -0.124  0.90144
CenterMA              8.40401    3.29003   2.554  0.01099 *
CenterMI             -5.55623    2.73766  -2.030  0.04303 *
CenterUC              3.85876    2.81022   1.373  0.17045
CenterWV              9.80935    3.02600   3.242  0.00128 **
Term02-2              0.14731    2.99520   0.049  0.96080
Term02-3              4.48264    2.36602   1.895  0.05883 .
Term03-1             -4.16629    2.10617  -1.978  0.04857 *
Term03-2             -0.66260    5.24515  -0.126  0.89953
Term03-3             -1.66283    2.22777  -0.746  0.45584
GenderMale            7.70661    1.64962   4.672 4.03e-06 ***
DBP                   0.19416    0.08685   2.236  0.02590 *
VLDL_TG              -0.16440    0.07770  -2.116  0.03494 *
RaceAsian            -4.67856    4.10674  -1.139  0.25525
RaceCaucasian        -1.23023    3.30170  -0.373  0.70963
RaceHispanic         -7.00634    4.50751  -1.554  0.12085
RaceOther             3.01057    4.90189   0.614  0.53944
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.23 on 419 degrees of freedom
  (192 observations deleted due to missingness)
Multiple R-squared:  0.2163,    Adjusted R-squared:  0.1789
F-statistic: 5.781 on 20 and 419 DF,  p-value: 2.177e-13
```

Validación modelo final

Se quiere asegurar que sea un buen modelo, por lo que se procederá a su validación mediante los residuos, que deberían ser normales y homocedásticos.

Se realizará un gráfico de los residuos vs los valores ajustados. La nube de puntos debería ser aleatoria y estar centrada en 0. Además, se realizará segundo gráfico, un Q-Q Plot para ver si se ajustan adecuadamente a la distribución normal (Figura 44).

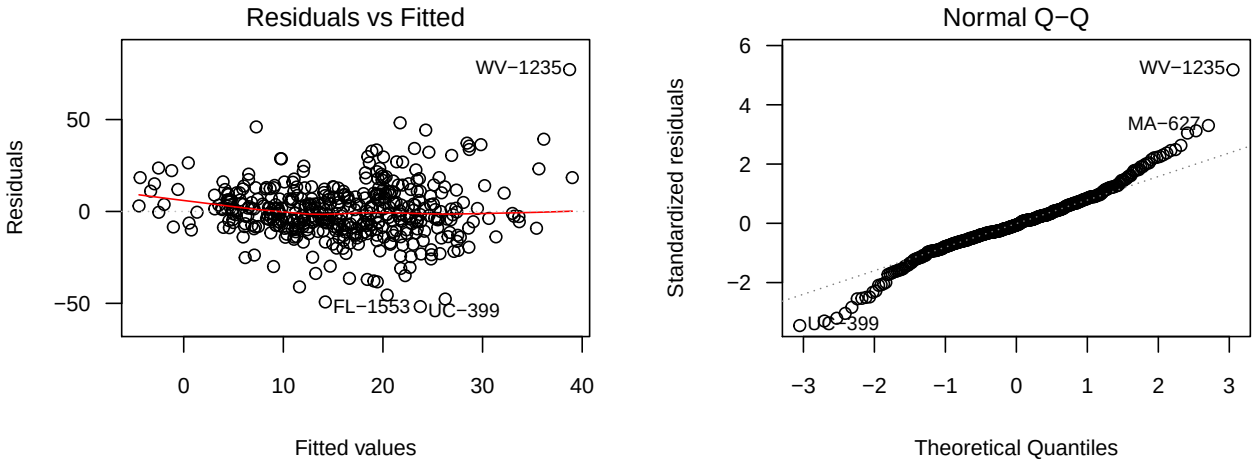

Figura 44: Gráfico residuos vs valores ajustados y QQ-Plot de los residuos del modelo codominante final para la ganancia de fuerza con las covariables y el SNP "vdr taq1" en el modelo codominante.

En el gráfico el primer gráfico, el de Residuos vs Valores Ajustados se puede ver una nube bastante aleatoria y parece centrada en el 0. Sin embargo, el punto del individuos "WV-1235" tiene un residuo muy superior al resto. En el Q-Q Plot parece que toda la parte del centro del gráfico se adecúa muy bien a la recta aunque las colas parecen distar de la misma un poco más. No obstante, no es suficiente para afirmar que los residuos no son normales. Lo que sí que llama más la atención es que el individuo "WV-1235" se aleja mucho de la normalidad. Este individuo no cumple ninguna de las dos hipótesis, por lo que se reestimará el modelo definitivo sin el mismo.

```
Call:
lm(formula = ND23_DIFF ~ as.factor(vdr_taq1) + Center + Term +
    Gender + DBP + VLDL_TG + Race, data = data2[-328, ])

Residuals:
    Min      1Q  Median      3Q     Max
-51.055  -7.985  -0.787   7.736  47.694

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -1.65633    8.11446  -0.204 0.838359
as.factor(vdr_taq1)CT  9.49923    2.36371   4.019 6.94e-05 ***
as.factor(vdr_taq1)TT  8.12500    2.42306   3.353 0.000872 ***
CenterFL             -2.81305    2.72726  -1.031 0.302923
CenterHH             -0.61083    3.32021  -0.184 0.854124
CenterMA              8.10775    3.18719   2.544 0.011323 *
CenterMI             -5.69687    2.65182  -2.148 0.032265 *
CenterUC              3.63619    2.72229   1.336 0.182371
CenterWV              8.14832    2.94738   2.765 0.005952 **
Term02-2              0.22020    2.90117   0.076 0.939535
Term02-3              3.33477    2.30174   1.449 0.148142
Term03-1             -4.12305    2.04004  -2.021 0.043910 *
Term03-2             -1.03113    5.08089  -0.203 0.839278
Term03-3             -1.80210    2.15797  -0.835 0.404143
GenderMale            7.19149    1.60072   4.493 9.11e-06 ***
DBP                   0.15735    0.08440   1.864 0.062981 .
VLDL_TG              -0.14789    0.07532  -1.963 0.050253 .
```

Modelo aditivo	Modelo recesivo	Modelo dominante	Modelo codominante
adrb2_1042713	vdr_tq1	resistin_g540a	vdr_tq1
resistin_g540a	ankrd6_q122e	igf1_t1245c	pcr5_snp1
kchj11_rs5219	pcr5_snp1	kchj11_rs5219	ankrd6_q122e
vdr_bsm1	vdr_bsm1	carp_c105t	vdr_bsm1
rs11630261	adrb2_1042713	akt1_c832g_c3359g	adrb2_1042713
gs_s287nga	vdr_rs731236	carp_a8470g	vdr_rs731236
akt1_c832g_c3359g	nr3c1_rs10482614	cast_rs754615	cast_rs7724759
resistin_c980g	cast_rs7724759	resistin_c980g	mgst3_4147542
akt1_g4362c	mgst3_4147542		fbox32_rs3739287
igf1_t1245c	fbox32_rs3739287		resistin_g540a
	akt1_c9756a_c11898t		nr3c1_rs10482614
	akt1_g4362c		tcfl72_7903146
	tcfl72_rs12255372		igf1_t1245c
	bcl6_4686467		kchj11_rs5219
			lepr_1137100
			fst_722910
			tcfl72_rs7903146

Tabla 6: SNP significativos (valor crítico 0.05) de los cuatro modelos para la ganancia de fuerza isométrica

```
RaceAsian             -4.59468    3.97780  -1.155 0.248718
RaceCaucasian         -1.45705    3.19830  -0.456 0.648935
RaceHispanic          -7.08516    4.36598  -1.623 0.105383
RaceOther              2.93982    4.74797   0.619 0.536138
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.75 on 418 degrees of freedom
  (192 observations deleted due to missingness)
Multiple R-squared:  0.2019,    Adjusted R-squared:  0.1637
F-statistic: 5.286 on 20 and 418 DF,  p-value: 5.904e-12
```

El R^2_{adj} del modelo es de 0.1637, indica que el modelo explica un 16.37% de la variabilidad de la ganancia de fuerza isométrica de los sujetos. El p-valor asociado al estadístico F que contrasta la utilidad del modelo es de 5.904e-12, menor al valor crítico 0.05. Se puede rechazar la hipótesis nula y decir que existe relación entre la variable respuesta y los predictores, es un modelo útil. Parece los individuos que tienden a ganar más fuerza isométrica en media son los hombres y los sujetos examinados en *University of West Virginia* y en *University of Massachusetts*. En *Central Michigan University* parece que se tiende a ganar menos fuerza isométrica que en el resto de universidades y los examinados en el trimestre "03-1" que en el resto de trimestres. Por otro lado, los dos niveles de la variable tienen un coeficiente asociado significativo. Las personas que tienen el genotipo "CT" en el gen "vdr_taq1" (heterocigotos) son los que tienden a ganar más fuerza, en media 9.5 unidades más, seguidos de los que tienen el genotipo "TT" en el gen. Los que tienden a ganar menos fuerza son los que tienen "CC" en el genotipo de "vdr_taq1". Se asemeja a lo que se había visto en los métodos gráficos.

Conclusiones

En la Tabla 6 se muestran los polimorfismos significativos de los cuatro modelos, son los SNPs que más influyen en la ganancia de fuerza isométrica. Se puede observar que los polimorfismos "mgst3_4147542", "fbox32_rs3739287", "igf1_t1245c", "ankrd6_q122e", "pcr5_snp1" aparecen en dos de los modelos, todos son de genes distintos. Por otro lado, los SNP "adrb2_1042713" y "kchj11 rs5219" aparecen en 3 de los modelos.

En el gen "resistin" se tienen dos SNPs significativos, "resistin_g540a", que aparece en los tres modelos, y "resistin_c980g", que aparece en dos modelos. Del gen "vdr" son significativos tres polimorfismos distintos: "vdr_bsm1" en 3 de los modelos, "vdr_tq1" en dos modelos y "vdr_rs731236" un modelo. El gen "akt1" también tiene tres polimorfismos distintos en la tabla: "akt1_c832g_c3359g" en modelos "akt1_g4362c", en un modelo "akt1_c9756a_c11898t" y en un modelo. Por último, del gen "cast" también se tienen dos SNPs significativos: "cast_rs7724759" en dos modelos y "cast_rs754615" en un modelo. Sin duda, el polimorfismo que más influye en la ganancia de fuerza isométrica es "vdr_tq1", el único con un p-valor asociado por debajo del nivel de significación de Bonferroni.

Aunque no era el objetivo principal del estudio también se ha visto qué covariables son las que globalmente son más importantes para la ganancia de fuerza isométrica. Se ha visto que los hombres y las personas con presión sanguínea diastólica alta son los que tienden a ganar más fuerza. Las personas que han sido estudiadas *University of West Virginia* y *University of Massachusetts* en media tienden a ganar más fuerza mientras que las estudiadas en *Central Michigan University* menos. Las a personas estudiadas en el trimestre "03-1" y con una cantidad de lipoproteínas de muy baja densidad alta les cuesta más ganar fuerza.

8.2 Ganancia en las pruebas repetición máxima (`NDRM_DIFF`)

Selección del modelo

Primeramente se considerará un modelo con la variable respuesta `NDRM_DIFF` y las variables `Center`, `Term`, `Gender`, `Age`, `Race`, `Pre.weight`, `Pre.height`, `pre.BMI`, `SBP`, `DBP`, `FLGU`, `TG`, `CHOL`, `HDL_C`, `CHOL_HDL_C`, `VLDL_TG`, `LDL_C`, `FINS`, `CRP`, `HOMA`, `Met_syn` como explicativas. Se procederá de la misma manera que para la variable `ND23_DIFF`.

Seguidamente, se utilizará la función `stepAIC` de la versión 7.3-48 del paquete `MASS` para realizar una selección de variables adecuada según el Criterio de Información de Akaike.

No se tendrán en cuenta las interacciones de primer orden en el modelo final ya que tras incluirlas se ha visto que no son estadísticamente significativas.

El test de que contrasta si el coeficiente asociado al parámetro de cada variable es o no distinto de 0 será:

$$\begin{cases} H_0 : \beta = 0 \\ H_1 : \beta \neq 0 \end{cases}$$

Los resultados del test se mostrarán a continuación, se realiza un test para cada coeficiente de cada variable. El nivel de significación que se utilizará es 0.05, por lo tanto, las variables que tengan un p-valor asociado al coeficiente de la misma inferior a este valor serán significativas y deberán incluirse, en principio, en el modelo.

```
Call:
lm(formula = NDRM_DIFF ~ Center + Term + Gender + Age + pre.BMI +
    DBP + HOMA + Race, data = lm1$model)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6762 -2.2625 -0.1909  2.1999  9.9054

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.28344    1.90602   4.346 1.73e-05 ***
CenterFL      3.21164    0.69743   4.605 5.44e-06 ***
CenterHH      1.96479    0.79175   2.482  0.01346 *
CenterIR      4.30645    3.59133   1.199  0.23114
```

```
CenterMA       2.13495    0.75604   2.824  0.00496 **
CenterMI       3.99708    0.65381   6.113 2.19e-09 ***
CenterUC       1.47720    0.63807   2.315  0.02108 *
CenterWV       3.16050    0.67093   4.711 3.34e-06 ***
Term02-2      -0.71772    0.70219  -1.022  0.30730
Term02-3       0.42313    0.52408   0.807  0.41989
Term03-1       1.11230    0.47550   2.339  0.01978 *
Term03-2      -0.65432    0.96349  -0.679  0.49743
Term03-3       0.20994    0.50219   0.418  0.67612
GenderMale     1.54761    0.36577   4.231 2.84e-05 ***
Age           -0.07281    0.03245  -2.244  0.02535 *
pre.BMI        0.10581    0.04074   2.597  0.00972 **
DBP           -0.03595    0.02075  -1.733  0.08384 .
HOMA          -0.17619    0.10995  -1.602  0.10981
RaceAsian     -1.79747    0.93454  -1.923  0.05509 .
RaceCaucasian -0.96921    0.75893  -1.277  0.20226
RaceHispanic  -0.76094    1.03987  -0.732  0.46471
RaceOther     -1.02177    1.14516  -0.892  0.37276
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.468 on 431 degrees of freedom
Multiple R-squared:  0.2203,    Adjusted R-squared:  0.1823
F-statistic: 5.798 on 21 and 431 DF,  p-value: 5.697e-14
```

Se vuelve a dudar de si las variables `Center` y `Term` deben incluirse en el modelo final, solamente son significativos algunos de las mismas. Consecuentemente se realizará una prueba F con la función `anova` (paquete `stats` versión 3.4.3) para comparar el modelo con cada una de las variables y sin ellas para comprobar si son distintos.

$$\begin{cases} H_0 : modelo\ nulo = modelo\ ampliado \\ H_1 : modelo\ nulo \neq modelo\ ampliado \end{cases}$$

```
Analysis of Variance Table

Model 1: NDRM_DIFF ~ Center + Term + Gender + Age + pre.BMI + DBP + HOMA +
    Race
Model 2: NDRM_DIFF ~ Term + Gender + Age + pre.BMI + DBP + HOMA + Race
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    449 5459.2
2    456 6136.7 -7   -677.47 7.9599 4.037e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Analysis of Variance Table

Model 1: NDRM_DIFF ~ Center + Term + Gender + Age + pre.BMI + DBP + HOMA +
    Race
Model 2: NDRM_DIFF ~ Center + Gender + Age + pre.BMI + DBP + HOMA + Race
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    449 5459.2
2    454 5596.1 -5   -136.93 2.2524 0.0483 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El p-valor asociado a la primera prueba F es inferior a 0.05 ($4.037e-09$). Se rechaza H_0 que enunciaba la igualdad de modelos. De la misma manera, en la prueba donde se contrasta la variable `Term` se puede observar que el p-valor del test es 0.0483, inferior también al nivel de significación. Se tienen evidencias estadísticamente significativas para rechazar la hipótesis nula, se puede decir que los modelos son distintos.

Como conclusión, tanto la variable `Center` como la variable `Term` se mantendrán en el modelo. Que aparezcan la variables `Center` y `Term` en el modelo, en principio, no es bueno ya que se supone que las pruebas que se realizan a los individuos están estandarizadas y son las mismas entre centros y trimestres. Se vuelve a dudar de la efectividad de la estandarización, la inclusión de estas variables en el modelo solventará en gran parte el problema.

En este caso también se añade la varable `Race` al modelo aunque se observa que su coeficiente no es estadísticamente significativo. Se ha visto en los apartados anteriores, tanto en el Análisis de Componentes Principales y en el equilibrio de Hardy-Weinberg que se tienen poblaciones no homogéneas, por lo que se cree importante su inclusión en el modelo.

Finalmente, el modelo que se utilizará es el que tiene como variable respuesta `NDRM_DIFF` y como explicativas `Center` (centro en que se realizan las pruebas), `Term` (timestre en que se realizan los test), `Gender` (género del individuo), `Age` (edad del sujeto), `pre.BMI` (índice de masa corporal antes del entrenamiento del individuo), `DBP` (presión sanguínea diastólica), `HOMA` (índice que permite precisar un valor numérico expresivo de resistencia insulínica) y `Race` (raza del individuo). Parece que las variables que explican la variable respuesta son muy parecidas a las del modelo anterior.

El R^2_{adj} del modelo es 0.1823, por lo cual sabemos que el modelo explica el 18.23% de la variabilidad de la variable respuesta.

Como se ha explicado anteriormente, el estadístico F es un buen indicador de si existe relación entre el predictor y las variables respuesta. Cuanto más lejos esté el estadístico de 1 mejor será. No obstante, este estadístico también será más grande cuantas más variables explicativas se tengan y más pequeño cuantos más datos se tengan. Se contrasta si es igual que 1 teniendo en cuenta los grados de libertad que se tienen:

$$\begin{cases} H_0 : \textit{No hay relación entre la variable respuesta y las variables explicativas} \\ H_1 : \textit{Hay relación entre la variable respuesta y las variables explicativas} \end{cases}$$

El p-valor del contraste es muy cercano a 0 ($5.697e-14$), entonces se puede rechazar la hipótesis nula y decir que existe relación entre la variable respuesta y los predictores, es un modelo útil.

A continuación, se incluirá en el modelo cada uno de los SNPs como variable explicativa con tal de poder identificar qué polimorfismos influyen más en la ganancia de fuerza en la prueba de repetición máxima.

Se seguirá el mismo procedimiento que en el apartado anterior. Como existe la posibilidad de que este tipo de variables sigan cuatro esquemas distintos, se considerarán los cuatro y se verán cuales son los que más influyen en la ganancia de fuerza en el test de repetición máxima.

Modelo aditivo

En este modelo el SNP se considera como variable numérica con tres posibles valores: 0, 1 y 2. Se supone que los niveles se agregan de manera conjunta para modelar los datos. Se considera el modelo con `ND23_DIFF` como variable respuesta y `Center`, `Term`, `Gender`, `DBP`, `VLDL_TG`, `Race` y el SNP como predictores.

Mediante un Q-Q plot de los p-valores asociados al coeficiente de cada SNP para contrastar si este coeficiente es igual a 0. Deberían ajustarse a una distribución uniforme. Se grafican con $-log_{10}$ para podernos centrar más en la primera parte de los valores (Figura 45).

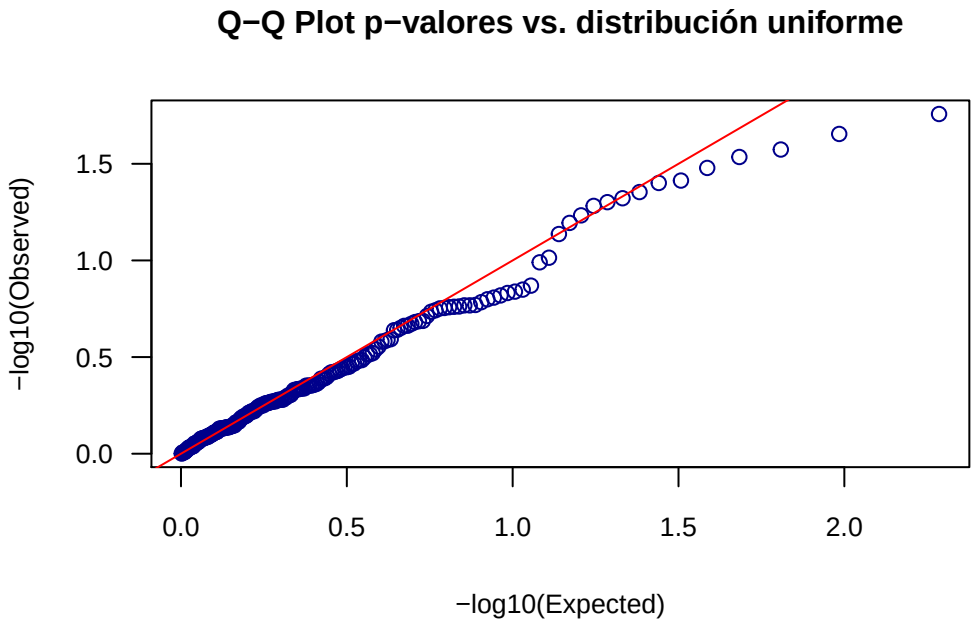

Figura 45: Q-Q Plot de los p-valores de todos los SNP para el modelo aditivo utilizando la ganancia del test de repetición máxima como variable respuesta.

Se observa que los p-valores se ajustan bastante bien a la recta roja en la cola de la izquierda, y por lo tanto, a una distribución uniforme. En cambio, en la cola de la derecha parecen distar un poco más, sin embargo no se tienen evidencias para decir que no siguen una distribución uniforme.

A continuación, se graficarán todos los p-valores asociados a los coeficientes del parámetro de cada polimorfismo. Se indicarán también dos rectas horizontales, la azul marcando el valor de crítico 0.05 y la roja con la Corrección de Bonferroni y los p-valores ajustados por el método FDR, para controlar la multiplicidad (Figura 46).

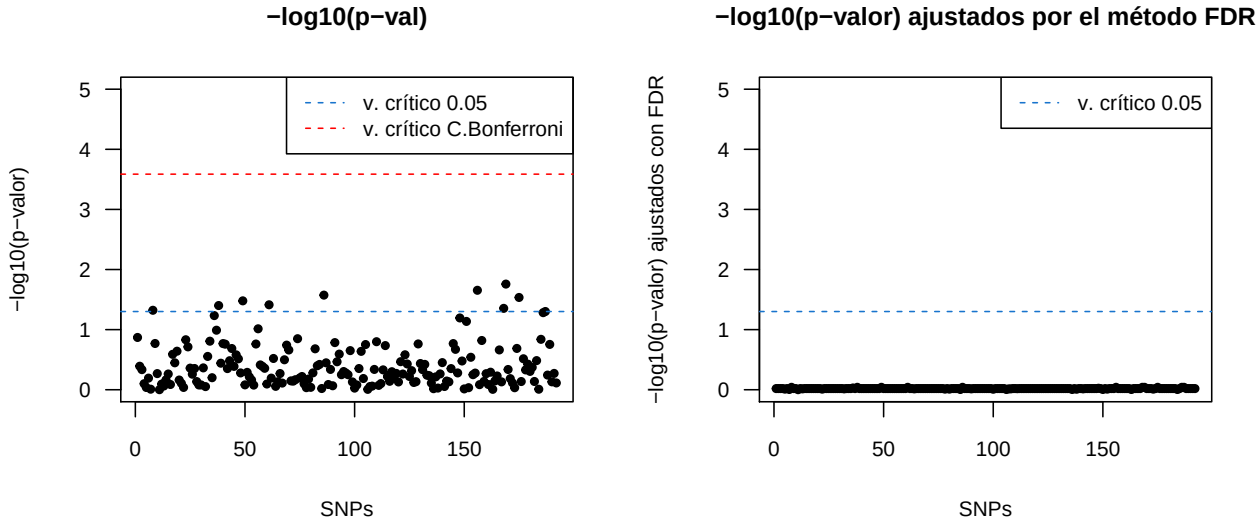

Figura 46: Gráfico de los p-valores de los SNP en el modelo aditivo para la ganancia en el test repetición máxima representando el v. crítico 0.05 y el v. crítico para la C. de Bonferroni y gráfico de los p-valores ajustados por el método FDR.

En el primer gráfico puede observarse que hay 10 polimorfismos por encima de la recta azul, con un p-valor

SNP	p-valor
resistin a537c	0.017
ppar gp12a	0.022
gapd 7971637	0.027
slc35f1 rs10484290	0.029
b2b	0.033
cav2 q130e	0.039
akt2 2304186	0.04
resistin c180g	0.044
adrb2 1042713	0.048
tpd52l1 3799736	0.05

Tabla 7: Lista de los diez polimorfismos con un p-valor más pequeño en el modelo aditivo para la ganancia del test repetición máxima.

asociado al coeficiente del SNP inferior a 0.05. Si se considerara este valor como nivel de significación, podría rechazarse la hipótesis nula que contrasta si el coeficiente es igual a 0. Sin embargo, no parece adecuado ya que al realizar 193 contrastes la probabilidad de obtener falsos positivos es alta. Como en los otros apartados se utiliza la Corrección de Bonferroni para controlar este error de tipo I. El nivel de significación con dicha corrección es $0.05/193 = 0.0002591$. Ninguno de los p-valores es inferior a esta cifra. Alternativamente, también se usa el método FDR pero tampoco da ningún p-valor significativo.

Se listan los diez polimorfismos con un p-valor más pequeño en la Tabla 7.

Puede observarse que el gen "resistin" aparece en dos ocasiones. Se estudiará más a fondo el SNP "resistin_a537c", el más significativo.

Se grafica el SNP respecto la variable respuesta para cada uno de sus niveles (Figura 47). Parece que el genotipo "AA" tiene una ganancia de fuerza en el test de repetición máxima superior al resto. Este SNP tiene 570 "AA" genotipos, 60 "AC" y 2 "CC".

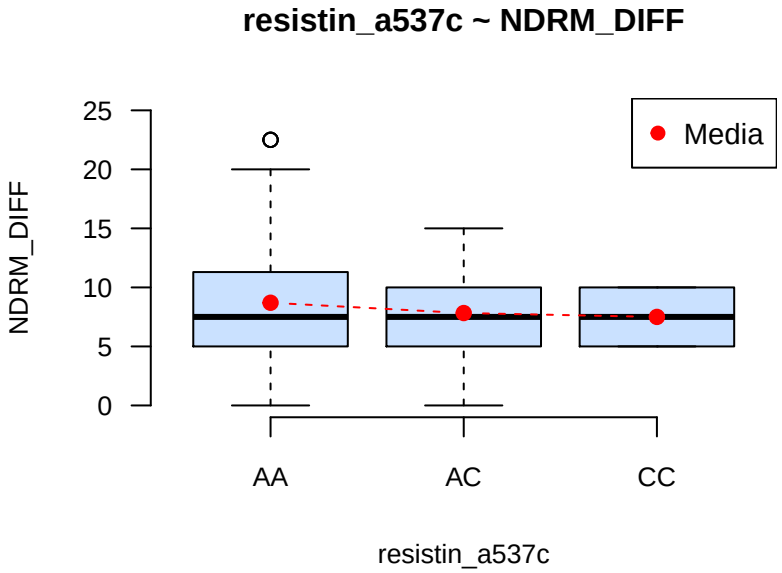

Figura 47: Test de repetición máxima según el SNP "resistin a537c" con la media

Se considera el modelo aditivo con `NDRM_DIFF` como variable respuesta y "resistin_a537c", `Center`, `Gender`, `DBP`, `Age`, `pre.BMI`, `HOMA` y `Race`, donde con "resistin_a537c" tiene 3 valores numéricos (0 para AA, 1 para

AC y 2 CC).

```
Call:
lm(formula = NDRM_DIFF ~ as.numeric(resistin_a537c) + Center +
    Gender + Age + pre.BMI + DBP + Race + HOMA, data = data2)

Residuals:
    Min      1Q  Median      3Q     Max
-9.2597 -2.4046 -0.1961  2.0242 10.8793

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               11.13932    2.01758   5.521 5.75e-08 ***
as.numeric(resistin_a537c) -1.26328    0.51410  -2.457  0.01438 *
CenterFL                   3.08075    0.69320   4.444 1.12e-05 ***
CenterHH                   1.92004    0.79189   2.425  0.01572 *
CenterIR                   3.12117    3.51108   0.889  0.37451
CenterMA                   2.26429    0.73217   3.093  0.00211 **
CenterMI                   4.19219    0.62691   6.687 6.89e-11 ***
CenterUC                   1.54382    0.62905   2.454  0.01451 *
CenterWV                   2.98080    0.66719   4.468 1.01e-05 ***
GenderMale                 1.78539    0.35948   4.967 9.75e-07 ***
Age                       -0.08214    0.03184  -2.579  0.01022 *
pre.BMI                    0.10787    0.04036   2.673  0.00781 **
DBP                       -0.04527    0.02039  -2.220  0.02690 *
RaceAsian                 -2.33706    0.91871  -2.544  0.01130 *
RaceCaucasian             -1.31055    0.75843  -1.728  0.08469 .
RaceHispanic              -1.45059    1.03555  -1.401  0.16198
RaceOther                 -1.38502    1.13642  -1.219  0.22359
HOMA                      -0.17876    0.10960  -1.631  0.10360
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.445 on 442 degrees of freedom
  (172 observations deleted due to missingness)
Multiple R-squared:  0.2354,    Adjusted R-squared:  0.206
F-statistic: 8.006 on 17 and 442 DF,  p-value: < 2.2e-16
```

El R^2_{adj} del modelo es de 0.206, cosa que indica que el modelo explica un 20.6% de la variabilidad de la ganancia de fuerza de los individuos. El p-valor asociado al estadístico F que contrasta la utilidad del modelo es casi 0 (< 2.2e-16), más pequeño que el valor crítico 0.05. Se puede rechazar la hipótesis nula y decir que existe relación entre la variable respuesta y los predictores, es un modelo útil. Parece ser que los hombres, las personas de menos edad, con una presión distólica baja y con un índice de masa corporal al inicio del estudio alto son las que tienden a ganar más fuerza después del entrenamiento en el test de repetición máxima. Parece que a los individuos de raza asiática les cuesta más ganar fuerza que al resto. Las personas que han sido estudiadas en los centros *University of Central Florida*, *University of Massachusetts*, *Central Michigan University*, *University of Connecticut* y *University of West Virginia* también parece que tienden a ganar más fuerza en este test, como se ha dicho anteriormente, parece que la uniformización entre centros no se ha llevado a cabo correctamente. Por último, en cuanto al SNP "resistin_a537c", parece ser que los individuos con el genotipo "CC" son los que tienden a ganar menos fuerza, seguidos de los que tienen el genotipo "AC". Los sujetos con el genotipo "AA" para este gen son los que tienden a ganar más fuerza para el test de repetición máxima. En media, si se aumenta en una unidad "C", la ganancia de fuerza en el test disminuye en 1.26 unidades. Este resultado concuerda con lo que se veía en la Figura 47.

Validación del modelo final

Para asegurarnos de que es un buen modelo, procederemos a su validación mediante los residuos, que deberían ser normales y homocedásticos.

En primer lugar, se graficarán los residuos vs los valores ajustados. Como se ha dicho anteriormente, la nube de puntos debería ser aleatoria y estar centrada en 0. Seguidamente, se mostrará un Q-Q Plot para ver si se ajustan adecuadamente a la distribución normal (Figura 48).

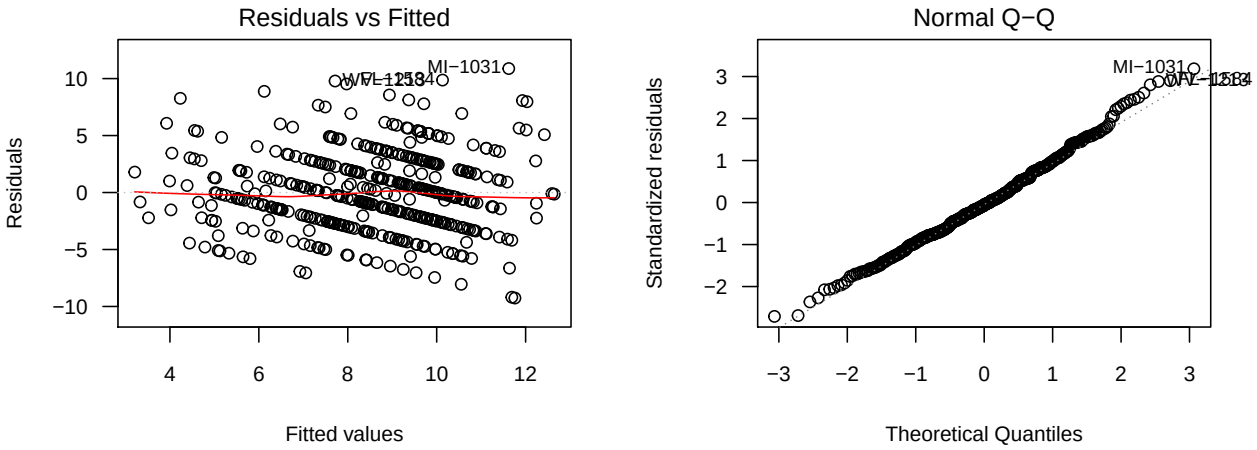

Figura 48: Gráfico residuos vs valores ajustados y QQ-Plot de los residuos del modelo aditivo final para la ganancia en el test de repetición máxima con las covariables y el SNP "resistin a537c" en el modelo.

En el gráfico de Residuos vs Valores Ajustados se puede observar una nube bastante aleatoria y centrada en 0. Parece que se forman rectas diagonales, es a causa de la variable respuesta solo tiene valores enteros. Parece que no hay ningún individuo con un residuo muy grande. En el Q-Q Plot parece que la distribución de los residuos se adecúa muy bien a la distribución normal. El modelo aditivo propuesto es adecuado.

Modelo recesivo

En este modelo el polimorfismo se considera como variable numérica con dos posibles valores: 0 para el homocigoto más frecuente y el heterocigoto y 1 para el homocigoto menos frecuente.

Se considera el modelo con recesivo `NDRM_DIFF` como variable respuesta y `Center`, `Gender`, `Age`, `pre.BMI`, `DBP`, `Race`, `HOMA` y el SNP, con los dos valores explicados anteriormente.

Se han eliminado de este modelo 11 SNPs que tienen solamente dos genotipos distintos, uno para los homocigotos y otro para los heterocigotos más comunes y el modelo recesivo solamente crea un nivel para estas dos categorías. El estudio de estos polimorfismos no aporta nada ya que no tienen variabilidad en este modelo, tienen el mismo nivel del factor para todos los individuos.

Se vuelve a realizar un Q-Q plot de los p-valores asociados al coeficiente de cada SNP que contrastan si este coeficiente es igual a 0. Deberían ajustarse a una distribución uniforme si se cumple H_0(Figura 49).

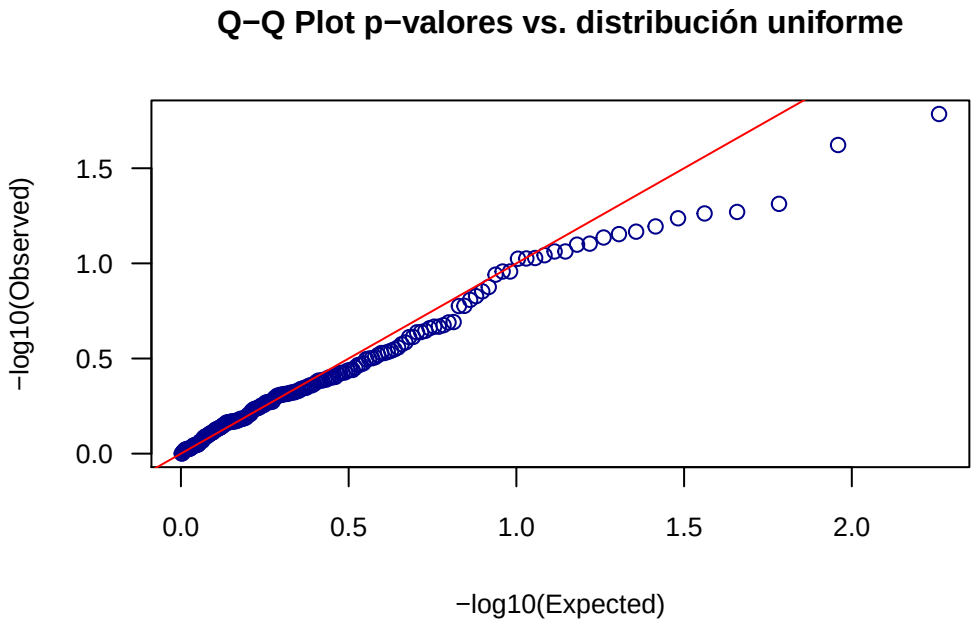

Figura 49: Q-Q Plot de los p-valores de todos los SNP del modelo recesivo para la ganancia de fuerza en el test de repetición máxima

La cola de la izquierda se ajusta bastante bien a una distribución uniforme como observarse. En cambio, en la cola de la derecha parece haber una serie de valores que distan de esta distribución.

Seguidamente, se realizará un gráfico con todos los p-valores asociados a los coeficientes del parámetro de cada SNP (Figura 50). Se mostrarán también dos rectas horizontales que indicarán el valor de crítico 0.05 y 0.05/193 con la Corrección de Bonferroni y en un segundo gráfico los p-valores ajustados por el método FDR, con tal de evitar el falso positivo.

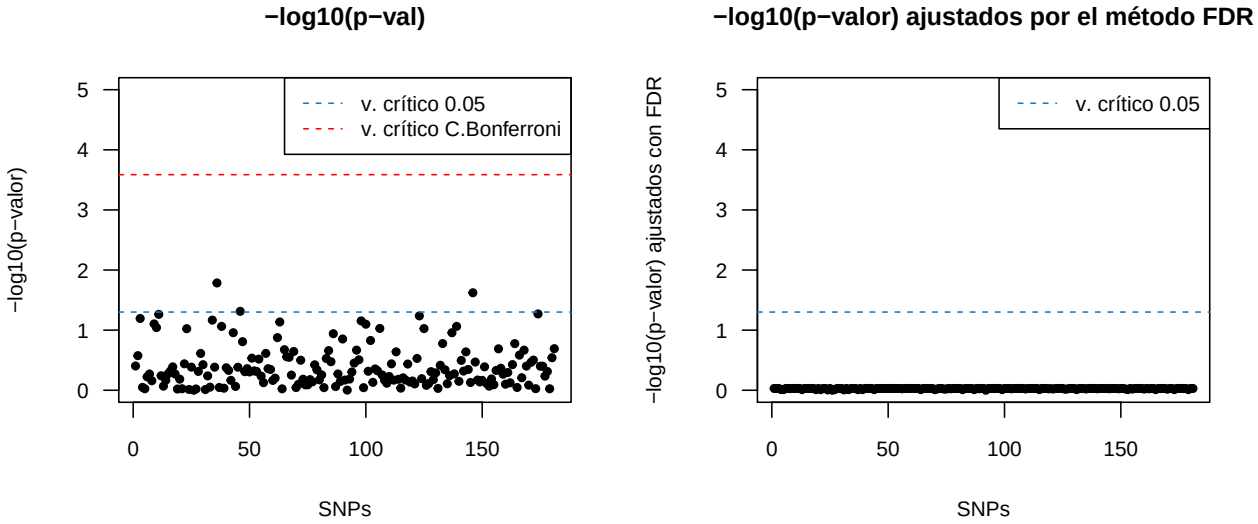

Figura 50: Gráfico de los p-valores de los SNP en el modelo recesivo para la ganancia en el test repetición máxima representando el v. crítico 0.05 y el v. crítico para la C. de Bonferroni y gráfico de los p-valores ajustados por el método FDR.

No se han encontrado SNPs muy significativos, por encima de la línea roja que marca el valor crítico usando

SNP	p-valor
akt2 2304186	0.016
ppar gp12a	0.024
bcl6 3774298	0.049
tpd52l1 4896782	0.054
adrb3 4994	0.055
nos3 rs1799983	0.058
actn3 r577x	0.064
akt2 7254617	0.068
igfbp3 6670	0.07
esr1 rs1042717	0.073

Tabla 8: Lista de los diez polimorfismos con un p-valor más pequeño con el modelo recesivo para el test de ganancia de fuerza en el test de repetición máxima.

la Corrección de Bonferroni. Se tienen 3 SNPs más por encima del valor crítico 0.05, que parecen tener relación con la variable respuesta aunque también debe tenerse en cuenta que la probabilidad de falso positivo cuando se realizan 193 test es alta. El método FDR tampoco da p-valores significativos, se tiene la certeza de que hay un máximo de falsos positivos del 5% tampoco da p-valores significativos.

La lista de los diez polimorfismos con un p-valor más pequeño se encuentra en la Tabla 8.

Se estudiará más a fondo el más significativo y primero de la lista: "akt2_2304186".

Se representa gráficamente el polimorfismo "akt2_2304186" respecto a la variable respuesta en cada uno de sus niveles (Figura 51). Según este gráfico, parece que "GG" y "GT" tienden a ganar más fuerza en el test de repetición máxima. Este SNP tiene 201 genotipos "GG", 302 "GT" y 129 "TT".

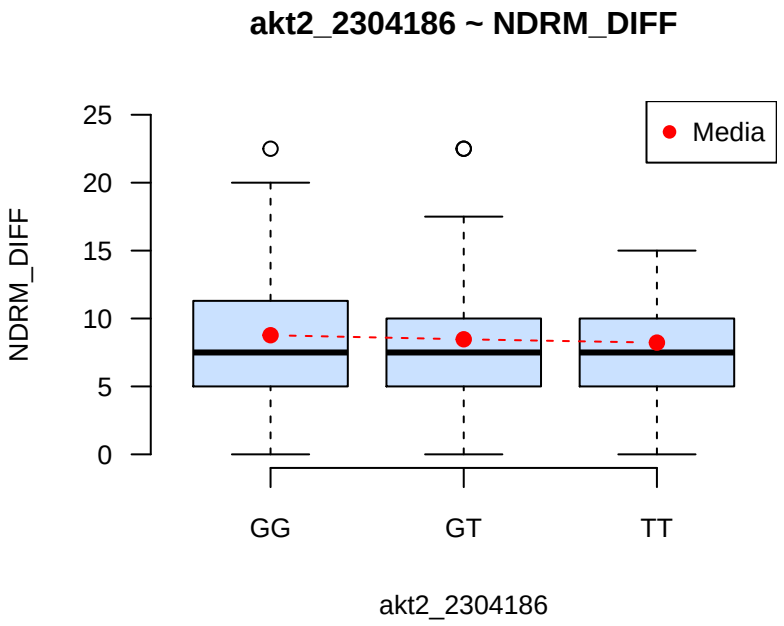

Figura 51: Test de ganancia de fuerza en el test de repetición máxima según el SNP "akt2 2304186" con la media.

Se considera el modelo con `NDRM_DIFF` como variable respuesta y "akt2_2304186", `Center`, `Gender`, `Age`, `pre.BMI`, `DBP`, `Race`, `HOMA` y el SNP como variables explicativas, donde "akt2_2304186" es una variable numérica con dos niveles. El nivel basal incluirá el nivel "GG" y "GT".

```
Call:
lm(formula = NDRM_DIFF ~ as.numeric(akt2_2304186b) + Center +
    Term + Gender + DBP + VLDL_TG + Race, data = data2)

Residuals:
   Min     1Q Median     3Q    Max
-8.875 -2.232 -0.150  2.110 10.865

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                7.712505   1.782618   4.327 1.88e-05 ***
as.numeric(akt2_2304186b) -1.044050   0.410146  -2.546 0.011255 *
CenterFL                   4.047164   0.651079   6.216 1.20e-09 ***
CenterHH                   2.222308   0.782107   2.841 0.004703 **
CenterIR                   5.051086   3.581207   1.410 0.159125
CenterMA                   2.728569   0.738149   3.697 0.000247 ***
CenterMI                   4.731805   0.624452   7.578 2.15e-13 ***
CenterUC                   1.978463   0.627383   3.154 0.001725 **
CenterWV                   3.475037   0.670219   5.185 3.32e-07 ***
Term02-2                  -1.063506   0.687346  -1.547 0.122529
Term02-3                   0.445096   0.522403   0.852 0.394675
Term03-1                   0.993575   0.474083   2.096 0.036680 *
Term03-2                  -0.480429   0.970297  -0.495 0.620755
Term03-3                   0.117729   0.502415   0.234 0.814843
GenderMale                 1.535296   0.367402   4.179 3.55e-05 ***
DBP                       -0.029365   0.019417  -1.512 0.131180
VLDL_TG                    0.001088   0.017269   0.063 0.949806
RaceAsian                 -1.466470   0.933267  -1.571 0.116835
RaceCaucasian             -0.545731   0.753036  -0.725 0.469022
RaceHispanic              -0.111313   1.033373  -0.108 0.914269
RaceOther                 -0.537046   1.149251  -0.467 0.640519
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.474 on 434 degrees of freedom
  (177 observations deleted due to missingness)
Multiple R-squared:  0.2129,    Adjusted R-squared:  0.1766
F-statistic: 5.869 on 20 and 434 DF,  p-value: 1.05e-13
```

El modelo explica un 17.66% de la variabilidad del modelo con las variables explicativas que se tienen ya que su R^2_{adj}. Como los modelos anteriores, el p-valor del `F-statistic` es casi 0 y por lo tanto significativo, cosa que quiere decir que los predictores ayudan a la explicación de la variable respuesta. Por otro lado, parece que los hombres, las personas que han sido estudiadas durante el trimestre "03-1" y en *University of Central Florida*, *Hartfort Hospital*, *Central Michigan University*, *University of Connecticut*, *University of Massachusetts* y *University of West Virginia*. Las personas con el alelo "TT" en el gen parece que tienden a ganar 1.04 unidades menos de fuerza que el resto, como se podía apreciar en el gráfico.

Validación modelo final

Se quiere asegurar que sea un buen modelo, por lo que se procederá a su validación mediante los residuos, que deberían ser normales y homocedásticos.

Primeramente, se graficarán los residuos vs los valores ajustados. La nube de puntos debería ser aleatoria y estar centrada en 0. Para continuar, se mostrará un Q-Q Plot para ver si se ajustan adecuadamente a la

distribución normal (Figura 52).

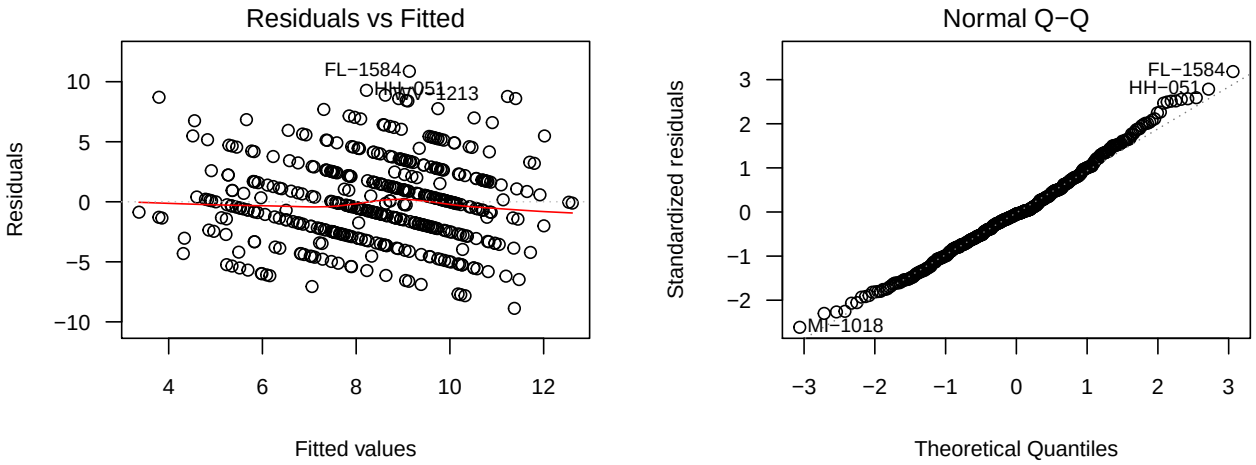

Figura 52: Gráfico residuos vs valores ajustados y QQ-Plot de los residuos del modelo recesivo final para la ganancia de fuerza con las covariables y el SNP "akt2 23041861" en el modelo recesivo.

No parece que haya grandes problemas en la validación de estos modelos. En en primer gráfico se observa una nube aleatoria y más o menos centrada en 0 como se esperaba. Por otro lado, en el Q-Q Plot sí que se observa que ambas colas distan un poco de la recta pero no es suficiente para decir que los residuos no son normales. Se valida el modelo correctamente.

Modelo dominante

En este modelo el SNP se considera como variable numérica con dos posibles valores: 0 para el homocigoto más frecuente y 1 para heterocigoto y el homocigoto menos frecuente. Se considera el modelo con dominante `NDRM_DIFF` como variable respuesta y `Center`, `Gender`, `Age`, `pre.BMI`, `DBP`, `Race`, `HOMA` y el SNP, con los dos valores explicados anteriormente.

Se vuelve a realizar un Q-Q plot de los p-valores asociados al coeficiente de cada polimorfismo que contrastan si este coeficiente es igual o no a 0. Deberían ajustarse a una distribución uniforme si la hipótesis nula es cierta para todos los valores (Figura 53).

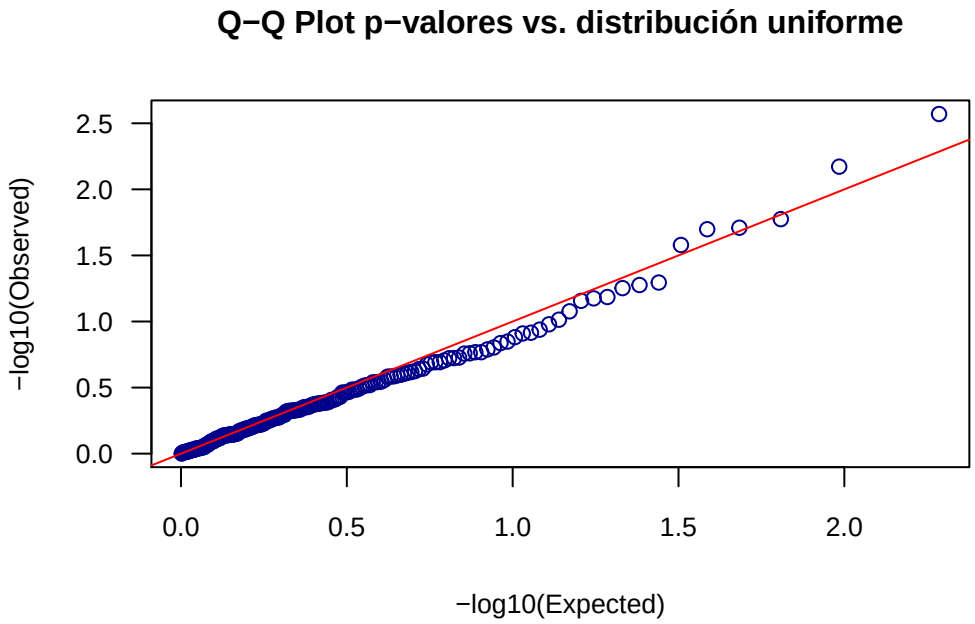

Figura 53: Q-Q Plot de los p-valores de todos los SNP del modelo dominante para la ganancia de fuerza del test de repetición máxima.

Los valores se acercan mucho a la distribución uniforme. Es probable que no haya p-valores extremadamente significativos para el modelo dominante.

Se realizará un gráfico con los p-valores asociados a los coeficientes del parámetro de cada SNP (Figura 54). Además, se mostrarán dos rectas horizontales que indicarán el valor de crítico 0.05 y 0.05/193 con la Corrección de Bonferroni para tratar de solventar el problema de multiplicidad.

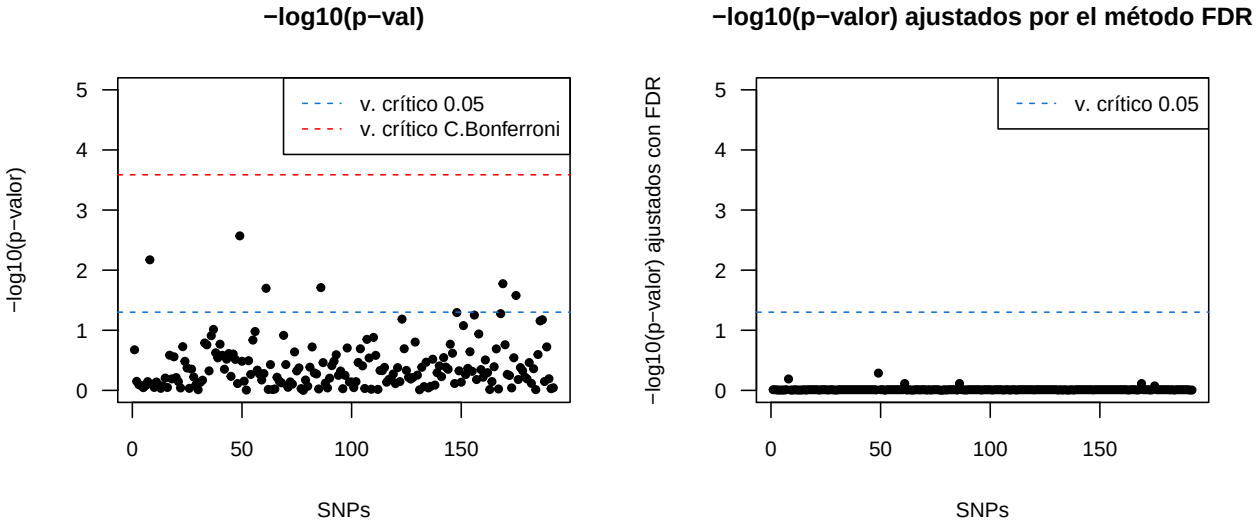

Figura 54: Gráfico de los p-valores de los SNP en el modelo dominante para la ganancia en el test repetición máxima representando el v. crítico 0.05 y el v. crítico para la C. de Bonferroni y gráfico de los p-valores ajustados por el método FDR.

Se tienen 6 valores por encima de la recta que marca el valor crítico 0.05 pero ninguno por encima de la que marca el valor crítico con la Corrección de Bonferroni ni con el método FDR que calcula la función `FDR` de la

SNP	p-valor
b2b	0.003
adrb2 1042713	0.007
resistin a537c	0.017
gapd 7971637	0.02
cav2 q130e	0.02
slc35f1 rs10484290	0.026
pcr15 snp4	0.051
resistin c180g	0.053
ppar gp12a	0.056
mgst3 4147542	0.065

Tabla 9: Lista de los diez polimorfismos con un p-valor más pequeño con el modelo dominante para la ganancia de fuerza del test de repetición máxima.

versión 1.9 del paquete `astsa`, se tiene la certeza de que se tienen, como máximo, un 5% de falsos positivos. También puede verse mediante el método gráfico.

Los diez polimorfismos más significativos se mostrarán en la Tabla 9

Se representa gráficamente el polimorfismo más significativo, "b2b", respecto a la variable respuesta en cada uno de sus niveles (Figura 55). Parece que el genotipo "TC" es el que tiende a ganar más fuerza en este test. Este SNP tiene 174 "CC" genotipos, 314 "TC" y 144 "TT".

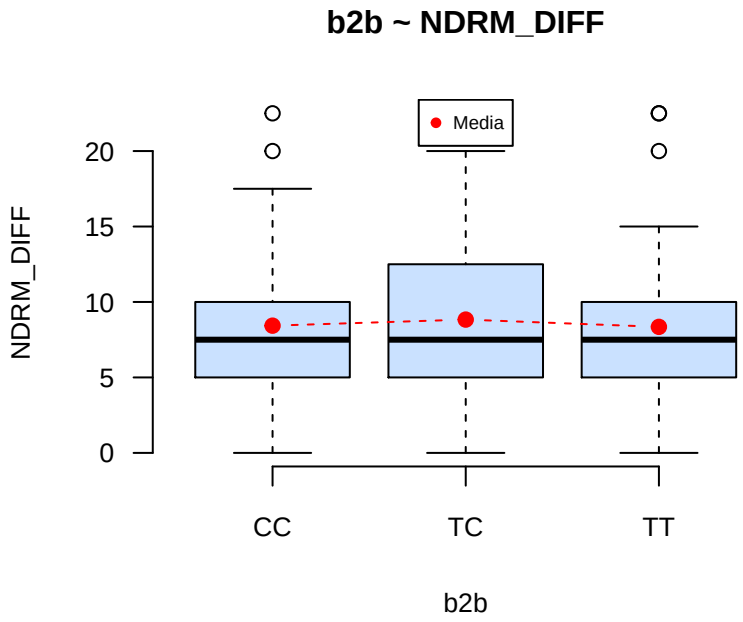

Figura 55: Test de repetición máxima de fu según el SNP "b2b" con la media.

Se considera el modelo con `NDRM_DIFF` como variable respuesta y "b2b", `Center`, `Gender`, `Age`, `pre.BMI`, `DBP`, `Race`, `HOMA` y el SNP como predictores, donde "b2b" es una variable numérica con dos niveles. El nivel basal incluirá el nivel "CC".

```
Call:
lm(formula = NDRM_DIFF ~ as.numeric(b2bc) + Age + Center + Term +
    Gender + DBP + Race + HOMA + pre.BMI, data = data2)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-9.3498 -2.2168 -0.2426  2.0800 11.7599

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       7.76946    1.88305   4.126 4.40e-05 ***
as.numeric(b2bc)  1.11047    0.36797   3.018  0.00269 **
Age              -0.07540    0.03212  -2.348  0.01933 *
CenterFL          3.06315    0.69356   4.417 1.26e-05 ***
CenterHH          1.87580    0.78858   2.379  0.01779 *
CenterIR          3.87142    3.57824   1.082  0.27986
CenterMA          2.04697    0.75104   2.726  0.00667 **
CenterMI          4.13078    0.62782   6.580 1.32e-10 ***
CenterUC          1.47179    0.63418   2.321  0.02075 *
CenterWV          3.18448    0.66742   4.771 2.48e-06 ***
Term02-2         -0.67370    0.69920  -0.964  0.33580
Term02-3          0.44476    0.50042   0.889  0.37460
Term03-1          1.27006    0.46824   2.712  0.00694 **
Term03-2         -0.69387    0.95898  -0.724  0.46972
Term03-3          0.43257    0.49936   0.866  0.38682
GenderMale        1.68871    0.36022   4.688 3.67e-06 ***
DBP              -0.03818    0.02034  -1.876  0.06124 .
RaceAsian        -2.40668    0.93499  -2.574  0.01037 *
RaceCaucasian    -1.13863    0.75899  -1.500  0.13427
RaceHispanic     -1.37214    1.03388  -1.327  0.18513
RaceOther        -1.42128    1.14674  -1.239  0.21585
HOMA             -0.16291    0.10962  -1.486  0.13796
pre.BMI           0.10871    0.04009   2.712  0.00695 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.456 on 448 degrees of freedom
  (161 observations deleted due to missingness)
Multiple R-squared:  0.2461,    Adjusted R-squared:  0.2091
F-statistic: 6.648 on 22 and 448 DF,  p-value: < 2.2e-16
```

El modelo explica un 20.91% de la variabilidad del modelo con los predictores que se tienen ya que su R^2_{adj}. Como los modelos anteriores, el p-valor del `F-statistic` es casi 0 y por lo tanto significativo, cosa que quiere decir que los predictores ayudan a la explicación de la variable respuesta. Por otro lado, parece que los hombres, los individuos de menos edad, los sujetos con un índice de masa corporal antes de empezar el estudio más alto, las personas que han sido estudiadas durante el trimestre "03-1" y en *University of Central Florida*, *Hartfort Hospital*, *Central Michigan University*, *University of Connecticut*, *University of Massachusetts* y *University of West Virginia*. Las personas con el genotipo "TC" y "TT" tienden a ganar 1.1 unidades más de fuerza en el test de repetición máxima como se veía en el gráfico.

Validación del modelo final

Se validan los residuos para ver si el modelo puede usarse. En primer lugar, se graficarán los residuos vs los valores ajustados. Como se ha dicho anteriormente, la nube de puntos debería ser aleatoria y estar centrada en 0. Seguidamente, se realiza un Q-Q Plot para ver si se ajustan adecuadamente a la distribución normal (Figura 56).

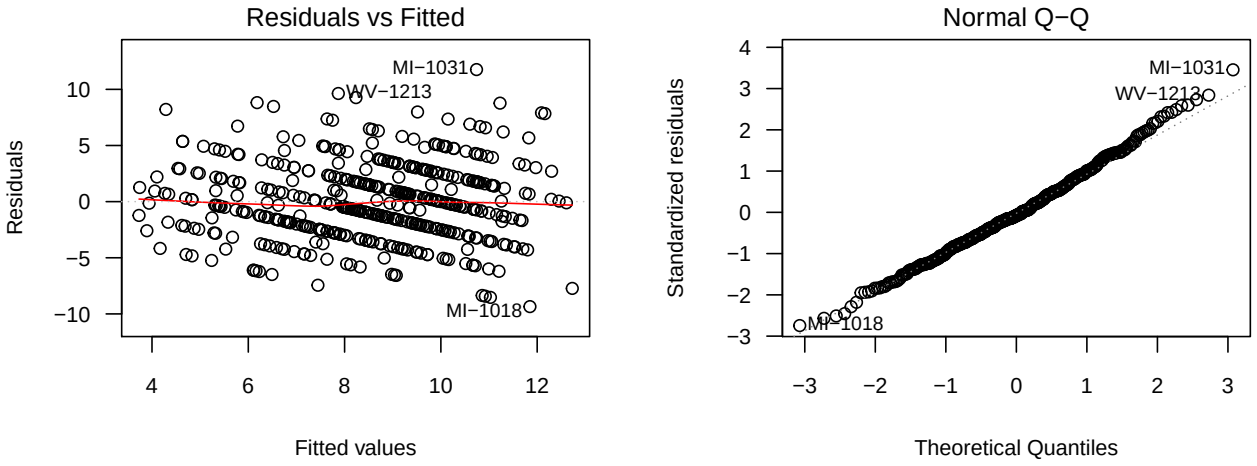

Figura 56: Gráfico residuos vs valores ajustados y QQ-Plot de los residuos del modelo dominante final para el test de ganancia de fuerza con las covariables y el SNP "b2b" en el modelo.

Se observan unas líneas diagonales a causa de que los valores de la variable respuesta son enteros. Parece no haber ningún problema de normalidad ni de hetrocedasticidad.

Modelo codominante

En este modelo se considerará la variable respuesta como una variable cualitativa, en la que se tendrán dos variables indicadoras: una para los heterocigotos y otra para el homocigoto menos común. El nivel basal serán los homocigotos más comunes. En R se las trata como factor.

Igual que en apartados anteriores, se realizará un Q-Q Plot para ver si los p-valores se asemejan a una distribución uniforme como se cree en la hipótesis nula (Figura 57).

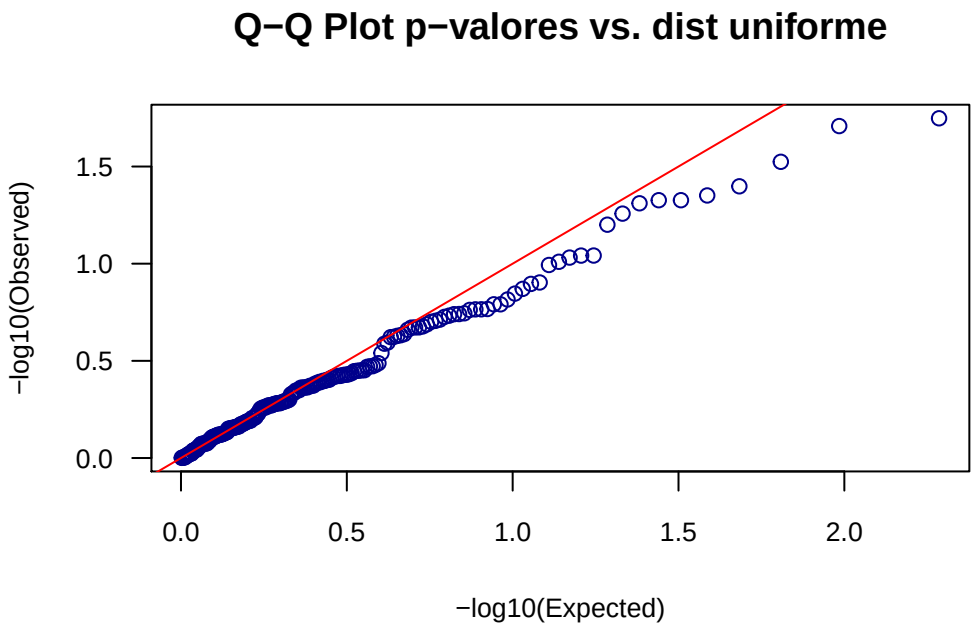

Figura 57: Q-Q Plot de los p-valores de todos los SNP para el modelo codominante utilizando la ganancia en el test de repetición máxima como variable respuesta.

SNP	p-valor
b2b	0.018
il15ra 3136618	0.02
adrb2 1042713	0.03
ppar gp12a	0.04
resistin a537c	0.045
ankrd6 a550t	0.047
pcr15 snp1	0.047
akt2 2304186	0.049
gapd 7971637	0.055
slc35f1 rs10484290	0.063

Tabla 10: Lista de los diez polimorfismos con un p-valor más pequeño en el modelo codominante para la ganancia del test de repetición máxima.

Es probable que no haya muchos p-valores significativamente pequeños ya que aunque se puede observar que se ajustan bastante bien al principio en los dos gráficos al final distan bastante por debajo de la recta.

A continuación, se realizará un gráfico con todos los p-valores asociados a los coeficientes del parámetro de cada polimorfismo para cada nivel del factor. De la misma manera que en el apartado anterior, se graficarán también dos rectas horizontales indicando el valor de crítico 0.05 y 0.05/193 con la Corrección de Bonferroni. También se realiza un segundo gráfico con los p-valores ajustados con el método FDR (Figura 58).

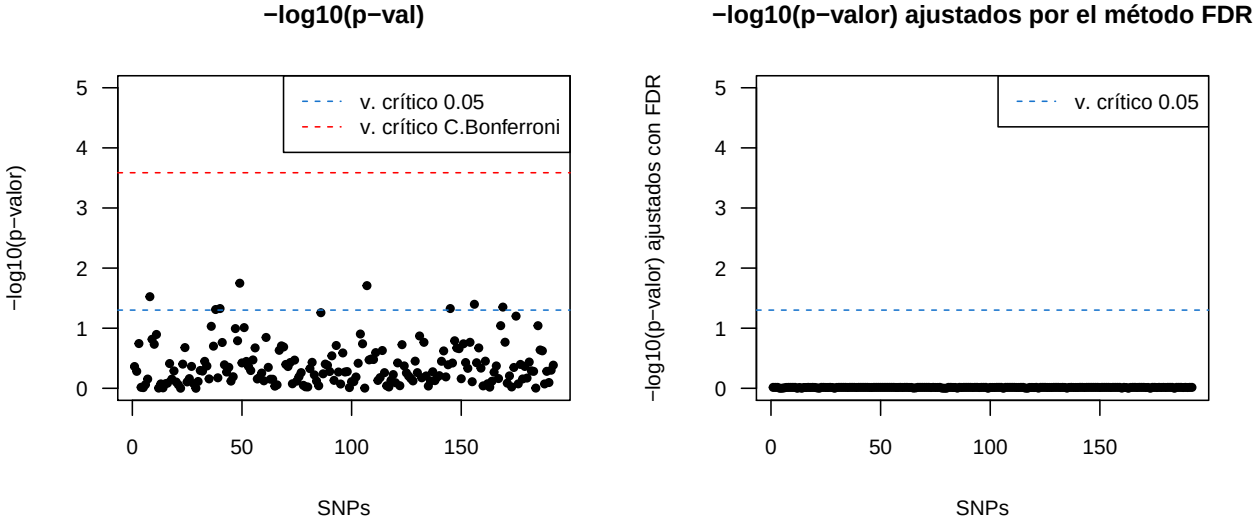

Figura 58: Gráfico de los p-valores de los SNP en el modelo codominante para la ganancia en el test repetición máxima representando el v. crítico 0.05 y el v. crítico para la C. de Bonferroni y gráfico de los p-valores ajustados por el método FDR.

En el primer gráfico puede observarse que hay 7 polimorfismos por encima de la recta azul, por lo tanto con un p-valor asociado al coeficiente del SNP inferior a 0.05. No hay ninguno por encima de la roja. Si consideráramos 0.05 este valor como nivel de significación, podría rechazarse la hipótesis nula que contrasta si el coeficiente es igual a 0 para estos 7 SNPs. Utilizando el criterio de Bonferroni y el *False Discovery Rate* ninguno resultaría importante para explicar la ganancia de fuerza en el test de repetición máxima.

La lista de los diez polimorfismos con un p-valor más pequeño se encuentra en Tabla 10.

A continuación, se considerará el polimorfismo "b2b", con el coeficiente asociado más significativo de los anteriores. Se grafica el SNP respecto la variable respuesta para cada uno de sus niveles (Figura 59). Parece

que el genotipo “TC” es el que tiende a ganar más fuerza en este test. Este SNP tiene 174 “CC” genotipos, 314 “TC” y 144 “TT”.

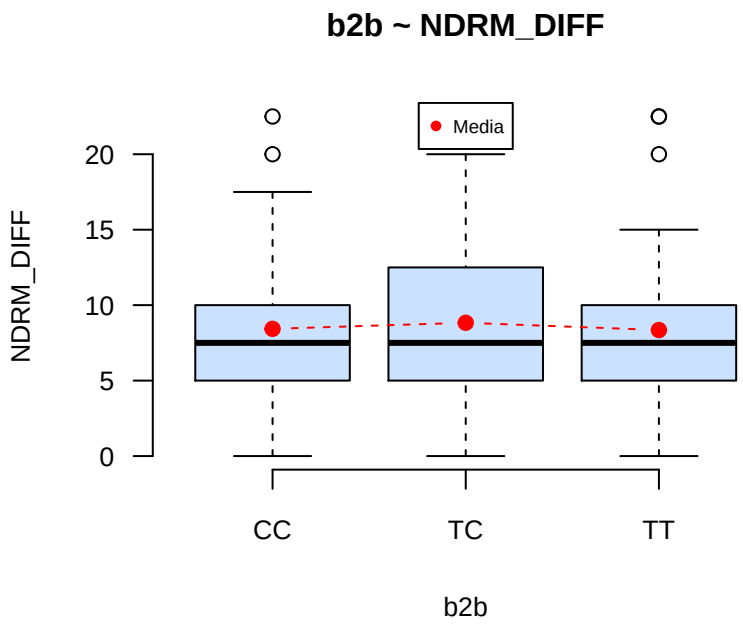

Figura 59: Test de repetición máxima según el SNP "b2b" con la media.

Se considera el modelo con `NDRM_DIFF` como variable respuesta y “b2b”, `Center`, `Gender`, `Age`, `pre.BMI`, `DBP`, `HOMA` y `Race`, donde con “b2b” es una variable factor con tres niveles, por lo que el modelo constará de dos variables indicadoras binarias. El nivel basal será el nivel del factor “CC”, y constará de una primera variable indicadora “TC” y para otra “TT”.

```
Call:
lm(formula = NDRM_DIFF ~ as.factor(b2bd) + Center + Gender +
    Age + pre.BMI + DBP + Race + HOMA, data = data2)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4202 -2.3551 -0.1736  2.2529 11.7282

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       8.58824    1.86418   4.607 5.32e-06 ***
as.factor(b2bd)TC 1.09592    0.38524   2.845  0.00465 **
as.factor(b2bd)TT 0.81360    0.49395   1.647  0.10023
CenterFL          2.99925    0.69411   4.321 1.91e-05 ***
CenterHH          1.69646    0.78521   2.161  0.03126 *
CenterIR          2.61190    3.55374   0.735  0.46274
CenterMA          2.22183    0.72783   3.053  0.00240 **
CenterMI          4.04145    0.62594   6.457 2.77e-10 ***
CenterUC          1.43329    0.62881   2.279  0.02311 *
CenterWV          3.02423    0.66981   4.515 8.09e-06 ***
GenderMale        1.74399    0.36273   4.808 2.08e-06 ***
Age              -0.08739    0.03194  -2.736  0.00647 **
pre.BMI           0.10157    0.04011   2.533  0.01166 *
```

```
DBP               -0.03578    0.02028  -1.764  0.07840 .
RaceAsian         -2.35983    0.95932  -2.460  0.01427 *
RaceCaucasian     -1.17704    0.76442  -1.540  0.12431
RaceHispanic      -1.28548    1.04060  -1.235  0.21735
RaceOther         -1.37101    1.15480  -1.187  0.23576
HOMA              -0.16053    0.11023  -1.456  0.14599
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.487 on 452 degrees of freedom
  (161 observations deleted due to missingness)
Multiple R-squared:  0.2254,    Adjusted R-squared:  0.1946
F-statistic: 7.307 on 18 and 452 DF,  p-value: < 2.2e-16
```

Las variables predictoras explican un 19.46% de la variabilidad del modelo. El p-valor asociado al estadístico F que contrasta si las variables explicativas en global ayudan a predecir la variable respuesta es de casi 0, es un modelo útil. Los hombres, las personas de menos edad, con una presión sistólica baja y con un índice de masa corporal al inicio del estudio alto y las personas de menos edad son las parece que tienden a ganar más fuerza después del entrenamiento en el test de repetición máxima. Parece que los individuos que han sido estudiados en los centros *University of Central Florida*, *University of Massachusetts*, *Central Michigan University*, *Hartfort Hospital*, *University of Connecticut* y *University of West Virginia* también parece que tienden a ganar más fuerza en el test, en este caso, también parece que la uniformización entre centros no se ha llevado a cabo de manera correcta. Se cree que a las personas de raza asiática les cuesta más ganar fuerza con el entrenamiento. Por último, en cuanto al SNP "b2b", parece ser que los individuos con el genotipo "TC" son los que tienden a ganar 1.09 unidades más de fuerza. Aunque el genotipo "TT" tiene un coeficiente positivo, no se puede decir que los sujetos con este haplotipo tiendan a ganar más fuerza en el test de repetición máxima ya que el p-valor asociado al coeficiente no es significativo. Estos resultados concuerdan con lo que se ve en el gráfico.

Validación del modelo final

Para asegurarnos de que es un buen modelo, procederemos a su validación mediante los residuos, que deberían ser normales y homocedásticos.

En primer lugar, se graficarán los residuos vs los valores ajustados. Como se ha dicho anteriormente, la nube de puntos debería ser aleatoria y estar centrada en 0. Seguidamente, se mostrará un Q-Q Plot para ver si se ajustan adecuadamente a la distribución normal (Figura 60).

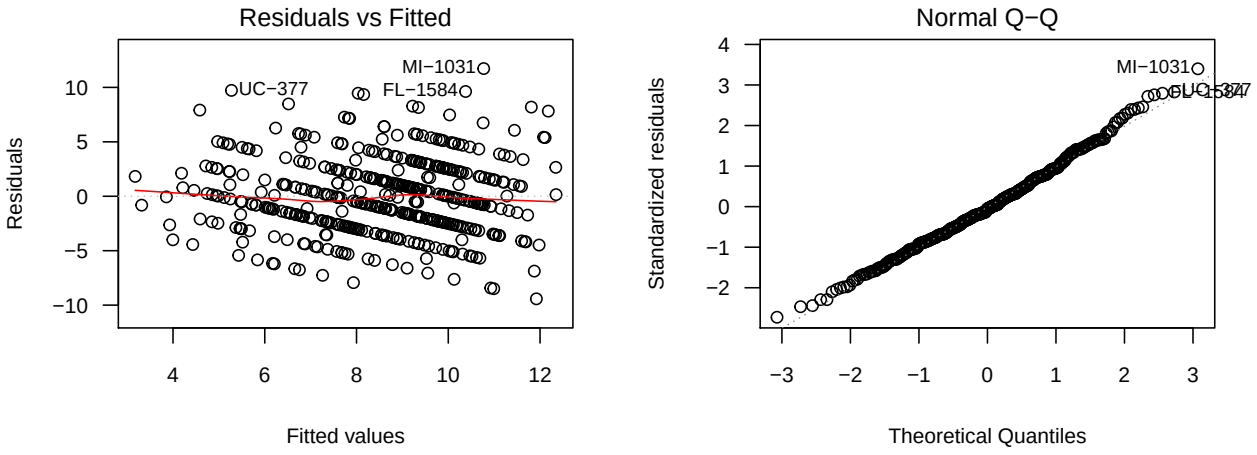

Figura 60: Gráfico residuos vs valores ajustados y QQ-Plot de los residuos del modelo codominante final para el test de repetición máxima con las covariables y el SNP "b2b" en el modelo.

Modelo aditivo	Modelo recesivo	Modelo dominante	Modelo codominante
resistin_a537c	akt2_2304186	b2b	b2b
ppar_gp12a	ppar_gp12a	adrb2_1042713	il15ra_3136618
gapd_7971637	bcl6_3774298	resistin_a537c	adrb2_1042713
slc35f1_rs10484290		gapd_7971637	ppar_gp12a
b2b		cav2_q130e	resistin_a537c
cav2_q130e		slc35f1_rs10484290	ankrd6_a550t
akt2_2304186			pcr15_snp1
resistin_c180g			akt2_2304186
adrb2_1042713			
tpd52l1_3799736			

Tabla 11: SNP significativos (valor crítico 0.05) de los cuatro modelos para el test de repetición máxima

En el primer gráfico se puede observar una nube bastante aleatoria y centrada en 0. Parece que se forman rectas diagonales, es a causa de la variable respuesta solo tiene valores enteros. Se puede observar que no hay ningún individuo con un residuo muy grande. En el Q-Q Plot parece que la distribución de los residuos se adecúa muy bien a la distribución normal.

Conclusiones

Se listan los polimorfismos significativos de los cuatro modelos en la Tabla 11. Son los SNPs que más contribuyen a la ganancia de fuerza en el test de repetición máxima. Los polimorfismos "cav2_q130e", "ppar_gp12a", "slc35f1_rs10484290" son significativos en dos de los modelos estudiados. Los que aparecen en tres de los modelos estudiados son: "ppar_gp12a", "b2b", "akt2_2304186" y "adrb2_1042713". Por otro lado, el gen "resistin" aparece en dos polimorfismos distintos: "resistin_a537c" que aparece en tres modelos y "resistin_c180g" que aparece en uno. El poimorfismo que más efecto tiene en la ganancia en el test de repetición máxima es "b2b".

Se ha visto que los hombres, las personas que tienen índice de masa corporal alto y las que tienen menos edad son las que tienden a ganar más fuerza en el test de repetición máxima. Se ha visto que las personas de raza asiática, en media, ganan menos fuerza en este test. En cuanto a los centros, en los que se tiende a ganar más fuerza son *University of Central Florida*, *University of Massachusetts*, *Central Michigan University*, *University of Connecticut*, *Hartfort Hospital* y *University of West Virginia*.

Se observa que el gen "resistin" aparecía en dos genes distintos para la variable respuesta de ganancia de fuerza isométrica en el polimofrismo "resistin_g540a", que aparecía en los tres modelos, y "resistin_c980g", que aparecía en dos. En la variable respuesta de ganancia en el test de repetición máxima "resistin_a537c" que aparece en tres modelos y "resistin_c180g" que aparece en uno. Entre las dos variables respuesta aparecen cuatro de los seis polimorfismos que se tienen de este gen, en el siguiente apartado se estudiará más a fondo con un tratamiento adecuado ya que se cree que es el que más influye en la ganancia de fuerza.

IX. Haplotipos del gen "resistin"

Anteriormente, se han realizado una serie de modelos estadísticos en los que se ha visto qué SNPs son los que más contribuyen a la ganancia de fuerza isométrica y en el test de repetición máxima. Para la realización de estos modelos se han tenido que hacer un número de test muy elevado, ya que los polimorfismos se han testeado uno por uno y la probabilidad de falso positivo, error de tipo I, es muy elevada. Para solventar este problema se han utilizado algunos métodos, como la Corrección de Bonferroni y el *False Discovery Rate*, sin embargo, en este apartado se propone otra alternativa. Se reducirá considerablemente el número de polimorfismos a testear creando nuevas variables, llamadas haplotipos, que serán combinaciones de SNPs del mismo gen.

Un haplotipo, técnicamente, es una combinación de alelos de *loci* adyacientes que son transmitidos juntos de generación en generación. Es sabido que los polimorfismos que se encuentran en el mismo cromosoma están muy correlacionados estadísticamente, se tratará de estudiar su asociación con las variables respuesta. El estudio de haplotipos puede ser biológicamente más relevante que el de los SNPs, además, los tests realizados suelen tener más potencia.

Este procedimiento simplifica el estudio y controla el problema de la multiplicidad, sin embargo también añade algo de incertidumbre a la base de datos ya que los haplotipos no se observan directamente. Hay muchos métodos estadísticos para la estimación de haplotipos, los más importantes son los métodos basados en la verosimilitud y los métodos bayesianos (Foulkes, 2004). En este estudio se utilizarán los primeros.

En el apartado anterior, se ha visto que cuatro de los seis polimorfismos del gen "resistin" eran significativos en las dos variables respuesta y en los dos modelos propuestos para cada una de ellas. Por otro lado, más del 79% de los individuos estudiados son de raza caucásica. Se ha visto durante el estudio mediante diversos métodos que se tenía una población no homogénea. Por lo que se realizará un estudio más profundo del gen "resistin" con haplotipos para los individuos de la raza caucásica. No se tienen suficientes individuos del resto de razas para un estudio fiable mediante este método.

Estimación de los haplotipos

Mediante el algoritmo de la función `haplo.em` de la versión 1.7.9 del paquete `haplo.stats` de R se estimarán los haplotipos de los seis polimorfismos del gen "resistin" para los individuos de raza caucásica, esta función basa la estimación en la máxima verosimilitud.

A continuación, se mostrarán los resultados de la estimación:

```
================================================================================
                                Haplotypes
================================================================================
   snp1 snp2 snp3 snp4 snp5 snp6 hap.freq
1     C    C    A    C    C    A  0.01262
2     C    C    A    C    C    C  0.00104
3     C    C    A    C    G    A  0.06044
4     C    C    A    G    C    A  0.00663
5     C    C    A    G    G    C  0.00096
6     C    C    G    C    C    A  0.25689
7     C    C    G    C    G    A  0.01321
8     C    C    G    C    G    C  0.03482
9     C    C    G    G    C    A  0.36154
10    C    C    G    G    G    A  0.01022
11    C    C    G    G    G    C  0.00372
12    C    T    A    C    C    A  0.00260
13    C    T    A    C    G    A  0.16176
14    C    T    A    G    C    A  0.00114
```

```
15    C    T    A    G    G    A  0.02344
16    C    T    G    C    C    A  0.00731
17    C    T    G    C    G    A  0.01302
18    C    T    G    C    G    C  0.00180
19    C    T    G    G    C    A  0.01253
20    C    T    G    G    G    A  0.00220
21    T    C    A    C    C    A  0.00053
22    T    C    G    C    C    A  0.01056
23    T    T    A    G    G    A  0.00101
=============================================================================
                                  Details
=============================================================================
lnlike =  -1530.816
lr stat for no LD =  1224.357 , df =  16 , p-val =  0
```

Los resultados muestran los haplotipos observados en el gen *resistin* y sus frecuencias estimadas, seguidamente se puede observar el *lr stat for no LD*, que es el estadístico del test de razón de verosimilitud que contrasta el *lnlike* para las frecuencias de los haplotipos estimadas *versus* el *lnlike* bajo la hipótesis nula de que los alelos de todo el *loci* están en equilibrio de ligamento. Que el `p-val` sea 0 quiere decir que la frecuencia de asociación de alelos es mayor o menor que lo que se podría esperar si los *loci* fueran independientes y asociados aleatoriamente.

El algoritmo calcula 23 haplotipos posibles con estos datos y con un umbral de 0.001 de los 64 posibles teóricos.

Los haplotipos más comunes son: "CCGCCA", "CCGGCA" y "CTACGA" con frecuencias del 36%, 25% y 16% respectivamente. Cada individuo tiene dos haplotipos, uno de la madre y otro del padre, se asigna a cada sujeto la combinación de haplotipos más probable con las probabilidades `posterior` del summary del objeto `haplo.em`.

Elaboración del modelo

Se elaborarán distintos modelos para las variables respuestas `ND23_DIFF` (ganancia de fuerza isométrica) y `NDRM_DIFF` (ganancia de fuerza en el test de repetición máxima). Por último, se considerará el modelo multivariante con las dos variables respuesta conjuntamente. Se tienen 61 niveles de la variable de los diplotipos, en primer lugar, se elaborará un modelo con los 4 niveles más frecuentes y una categoría "otros" y, en segundo lugar, se elaborará otro modelo con todos los niveles, para ver cómo se comportan los más minoritarios. Un diplotipo consiste la combinación de dos alelos, uno materno y otro paterno.

Ganancia de fuerza isométrica (`ND23_DIFF`)

Para tener una visión general de esta variable para los distintos haplotipos, se realizará un *boxplot* para los cuatro diplotipos (combinaciones de dos haplotipos) más frecuentes (Figura 61) y otro para el resto. Estas cinco combinaciones de haplotipos tienen una frecuencia del 21.2%, 13.3%, 13%, 9.1% y 43.4% respectivamente. De esta manera, podrán compararse entre sí:

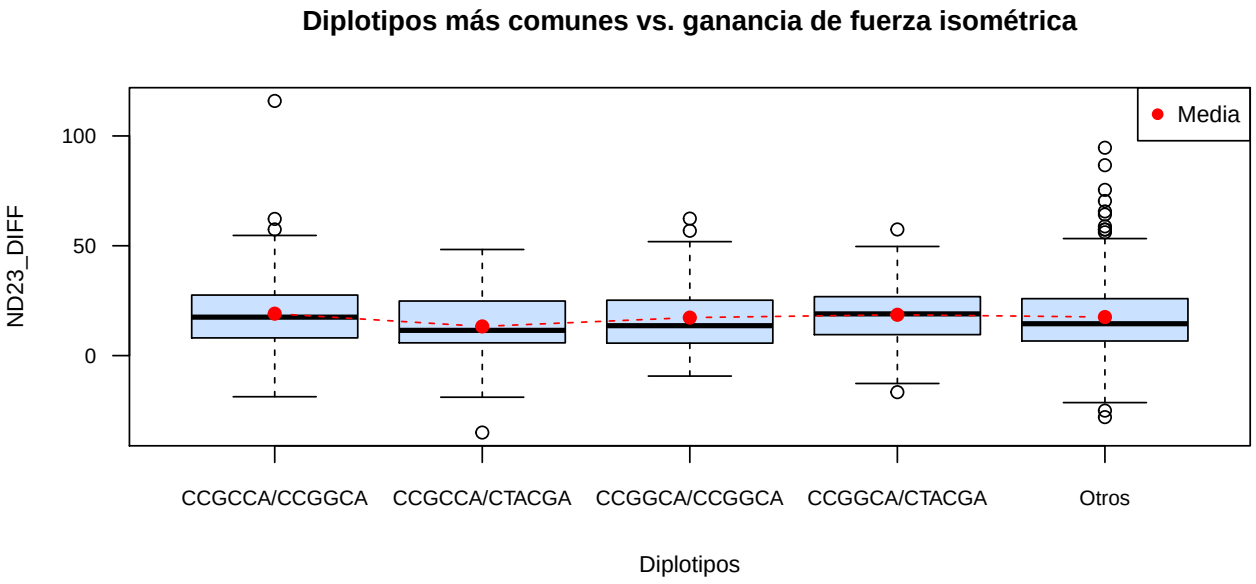

Figura 61: Boxplots de la ganancia de fuerza isométrica del brazo no dominante estratificada por diplotipos más comunes.

Puede observarse que la mediana no parece muy distinta entre distintos los grupos, no obstante, cabe destacar que en el grupo "Otros" parece haber más dispersión. Por otro lado, los sujetos del tercer grupo (homocigotos para el haplotipo "CCGGCA/CCGGCA") y los del cuarto (CCGGCA/CTACGA) parecen que tienen una mediana un poco superior al resto.

Seguidamente, se procederá a la elaboración de modelos lineales para analizar mejor estos fenómenos. Se utilizarán los cuatro diplotipos más frecuentes y la categoría "otros" junto con las covariables significativas (p-valor asociado al coeficiente del parámetro inferior a 0.05). Se contrastarán mediante la función ANOVA, ya explicada anteriormente.

```
Anova Table (Type III tests)

Response: ND23_DIFF_cau
                    Sum Sq  Df F value    Pr(>F)
(Intercept)             15   1  0.0639  0.800541
Center                9947   7  6.0735 1.096e-06 ***
Term                  3605   5  3.0811  0.009829 **
Gender                3695   1 15.7930 8.662e-05 ***
DBP                   1245   1  5.3204  0.021688 *
VLDL_TG               1137   1  4.8612  0.028149 *
as.factor(orvhapl2)    873   4  0.9333  0.444713
Residuals            78148 334
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Las covariables `Center` (centro en el que se realizan las pruebas), `Term` (trimestre en el que se estudia al individuo), `Gender` (género del sujeto), `DBP` (presión sanguínea diastólica) y `VLDL_TG` (lipoproteínas de muy baja densidad) sí que ayudan a explicar la ganancia de fuerza isométrica. En cambio, no se tienen evidencias de que los diplotipos más frecuentes del gen "resistin" influyan en esta ganancia.

```
Call:
lm(formula = ND23_DIFF_cau ~ Center + Term + Gender + DBP + VLDL_TG +
    as.factor(orvhapl2), data = dcau)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-49.624  -8.340  -0.325   7.781  77.089

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                        2.16540    8.56401   0.253  0.80054
CenterFL                          -1.60046    3.57686  -0.447  0.65484
CenterHH                          -0.87678    4.33668  -0.202  0.83990
CenterIR                          19.93602   15.94426   1.250  0.21204
CenterMA                           5.63544    3.96991   1.420  0.15667
CenterMI                          -5.94764    3.17547  -1.873  0.06194 .
CenterUC                           5.28638    3.33726   1.584  0.11413
CenterWV                           9.65924    3.40261   2.839  0.00481 **
Term02-2                          -1.86532    3.22318  -0.579  0.56317
Term02-3                           3.62261    2.64833   1.368  0.17227
Term03-1                          -6.42570    2.38438  -2.695  0.00740 **
Term03-2                          -4.74524    4.55146  -1.043  0.29790
Term03-3                          -1.89779    2.58488  -0.734  0.46335
GenderMale                         7.49075    1.88492   3.974 8.66e-05 ***
DBP                                0.23420    0.10153   2.307  0.02169 *
VLDL_TG                           -0.18848    0.08548  -2.205  0.02815 *
as.factor(orvhapl2)CCGCCA/CTACGA -5.91383    3.19945  -1.848  0.06543 .
as.factor(orvhapl2)CCGGCA/CCGGCA -1.05912    2.95791  -0.358  0.72052
as.factor(orvhapl2)CCGGCA/CTACGA -1.47989    2.94534  -0.502  0.61568
as.factor(orvhapl2)Otros         -2.34782    2.19465  -1.070  0.28549
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.3 on 334 degrees of freedom
  (142 observations deleted due to missingness)
Multiple R-squared:  0.2146,    Adjusted R-squared:  0.1699
F-statistic: 4.802 on 19 and 334 DF,  p-value: 6.433e-10
```

Con el `summary` de este modelo puede observarse que hay uno de los diplotipos, "CCGCCA/CTACGA", cuyo coeficiente roza la significación. Se estudiará, a continuación, la misma variable respuesta con todos los niveles de diplotipos para ver si estudiando la variable más a fondo pueden sacarse más conclusiones. También se quiere averiguar cómo se comportan los niveles más minoritarios. En este modelo con la ganancia de fuerza isométrica como variable respuesta, también se incluirán solamente las covariables con un p-valor asociado inferior a 0.05.

```
Anova Table (Type III tests)

Response: ND23_DIFF_cau
                   Sum Sq  Df F value    Pr(>F)
(Intercept)           135   1  0.5790   0.44715
Center              12007   7  7.3348 2.728e-08 ***
Term                 3348   5  2.8632   0.01488 *
Gender               2949   1 12.6083   0.00043 ***
DBP                  1709   1  7.3078   0.00716 **
as.factor(orvhapl)  20833  60  1.4847   0.01531 *
Residuals           93077 398
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Las covariables `Term`, `Gender`, `DBP` y `Center` son importantes en el modelo, el p-valor asociado al coeficiente del parámetro que contrasta si son iguales o no a 0 es significativo (inferior a 0.05). Además, el coeficiente del parámetro de la variable que contrasta los diplotipos también es significativa. Puede decirse que las combinaciones de haplotipos estimada del gen "resistin" influye en la ganancia de fuerza isométrica. Mediante la función `summary` de la versión 3.6.0 del paquete `base` de R se observa que la combinación de haplotipos con coeficiente significativo son: "CCGGCA/CTGGGA", "CCGGCA/TCGCCA", "CCGGGA/CCGGGA", "CCGGGA/CTACGA", "CTACGA/CTGCGA" y "CTACGA/TCGCCA". Se tienen, respectivamente 10, 19, 11, 27, 30 y 39 de 496 individuos caucásicos que se estudian por cada combinación de haplotipos, que corresponden al 2%, 3.8%, 2.2%, 5.4%, 6% y 7.8% del tamaño de la muestra. Se puede observar que el haplotipo "CTACGA" aparece en las tres últimas combinaciones, que son las más frecuentes. Por otro lado, el haplotipo "CCGGGA" aparece en un heterocigoto y forma de homocigoto pero se descarta su estudio exhaustivo ya que solamente lo posee un 7% de la muestra. Se cree que el haplotipo del gen "resistin" que más influye en la ganancia de fuerza isométrica es "CTACGA", lo posee un 32% de la muestra.

Se crea una nueva variable dicotómica con el nivel "0" si el individuo no posee el haplotipo "CTACGA" y "1" si sí que lo posee. Hay 339 individuos que no lo poseen y 157 que sí, representan el 68% y el 32% de la muestra respectivamente. Se vuelve a realizar el modelo con la variable respuesta ganancia de fuerza isométrica y las covariables y la nueva variable como predictoras. Mediante la función ANOVA se evaluará qué variables predictoras son las más importantes.

```
Anova Table (Type III tests)

Response: ND23_DIFF_cau
                Sum Sq  Df F value    Pr(>F)
(Intercept)         11   1  0.0468   0.82887
Center           10694   7  6.5883 2.623e-07 ***
Term              3963   5  3.4181   0.00501 **
Gender            4136   1 17.8356 3.102e-05 ***
DBP               1235   1  5.3270   0.02160 *
VLDL_TG           1075   1  4.6361   0.03201 *
as.factor(CTACGA)  874   1  3.7688   0.05305 .
Residuals        78147 337
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Las variables `Center`, `Term`, `Gender`, `DBP`, `VLDL_TG` salen significativas, cosa que ya se ha visto a lo largo del estudio Lo que hay que destacar, es que la variable que indica si un individuo posee el haplotipo "CTACGA" está rozando el nivel de significación 0.05. Utilizando un criterio un poco más laxo y poniendo el nivel de significación en 0.1 sí que podría decir que este haplotipo influye positivamente en la ganancia de fuerza. Por el momento, lo único que puede decirse es que se confirma que se tienen indicios de que este haplotipo del gen "resistin" está relacionado con la ganancia de fuerza isométrica en individuos de raza caucásica. Puede ser un inicio para, en otros estudios, estudiar este fenómeno específico. En el `summary` que se encuentra a continuación, puede verse que las personas con el haplotipo "CTACGA", en media, ganan 3.49 más unidades de fuerza isométrica que las que no lo tienen.

Ganancia de fuerza en el test de repetición máxima (`NDRM_DIFF`)

Se realizará el mismo procedimiento anteriormente para la ganancia de fuerza isométrica en el test de repetición máxima. Se procede a la realización de un *boxplot* para las cuatro combinaciones de haplotipos más frecuentes (Figura 62) y otro para el resto para poderlos comparar:

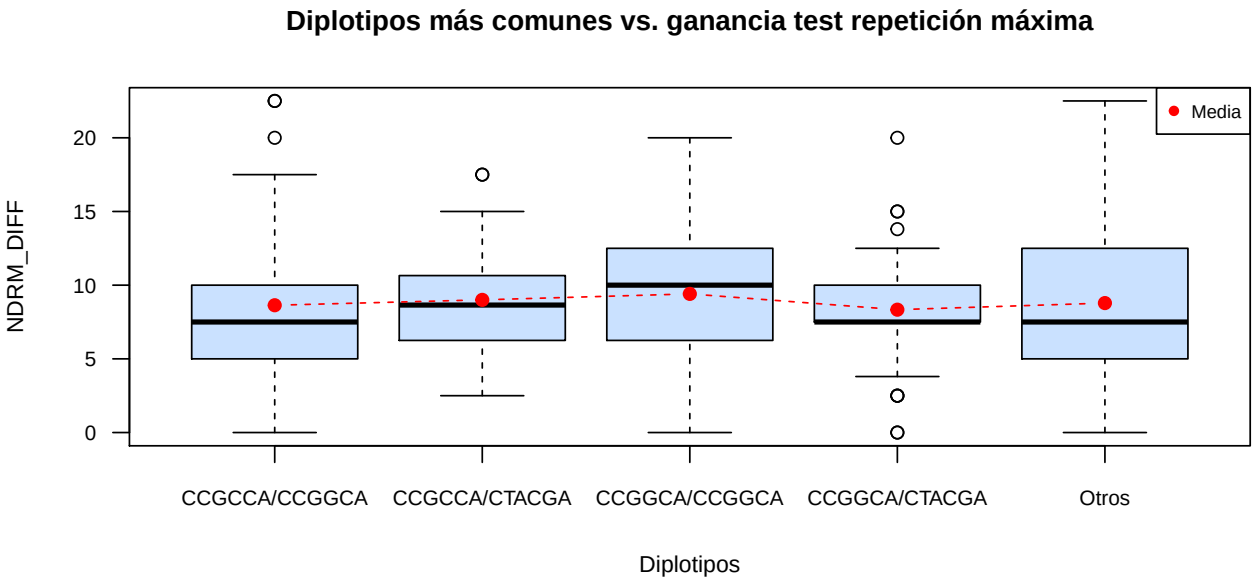

Figura 62: Boxplots de la ganancia en el test de repetición máxima del brazo no dominante estratificada por diplotipos más comunes.

Para esta variable parece haber más diferencias entre las medianas y los cuantiles de los cuatro grupos. El genotipo del tercer gráfico "CCGGCA/CTACGA" parece tener una mediana ligeramente superior al resto. Igual que para la variable de fuerza isométrica, parece que en el grupo de "Otros" parece haber más dispersión. Contrariamente, el haplotipo "CCGCCA/CTACGA" (cuarto *boxplot*) parece que es el que menos variabilidad hay, la mitad de individuos tienen una ganancia de fuerza en el test de repetición máxima de entre 7.5 y 10. Se estudiarán estas diferencias entre los diplotipos más frecuentes mediante el modelo lineales que se muestra a continuación. Se contrastará, igual que en el apartado anterior, con la función ANOVA.

```
Anova Table (Type III tests)

Response: NDRM_DIFF_cau
                    Sum Sq  Df F value    Pr(>F)
(Intercept)          208.5   1 15.1380 0.0001148 ***
Center               785.7   7  8.1486 2.301e-09 ***
Term                 255.5   5  3.7104 0.0026539 **
Gender               167.4   1 12.1509 0.0005379 ***
DBP                   81.0   1  5.8840 0.0156643 *
pre.BMI               70.1   1  5.0881 0.0245621 *
as.factor(orvhapl2)   39.8   4  0.7232 0.5763941
Residuals           6308.5 458
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El centro donde se realizan las pruebas, el trimestre en que se estudia al individuo, el género, la presión sanguínea diastólica y el índice de masa corporal antes de empezar el estudio son importantes a la hora de explicar la ganancia de fuerza en el test de repetición máxima. Contrariamente, no se tienen evidencias de que los diplotipos más frecuentes del gen "resistin" influyan en esta ganancia. Se ha realizado el `summary` del modelo y ningún nivel de esta variable se acerca al nivel de significación.

Igual que se ha hecho en el apartado anterior, se estudiarán los diplotipos con todos sus factores para ver también cómo se comportan los más minoritarios y si tienen relación con la ganancia de fuerza en el test de repetición máxima. Se realiza el modelo con las covariables y los diplotipos. Se contrastarán los resultados con la función ANOVA.

```
Anova Table (Type III tests)

Response: NDRM_DIFF_cau
                   Sum Sq  Df F value    Pr(>F)
(Intercept)          36.5   1  2.5980  0.107775
Center              478.5   7  4.8684 2.735e-05 ***
Gender              244.0   1 17.3798 3.743e-05 ***
DBP                  63.8   1  4.5438  0.033639 *
Age                 100.2   1  7.1358  0.007860 **
pre.BMI             111.5   1  7.9419  0.005067 **
as.factor(orvhapl)  812.2  59  0.9805  0.520165
Residuals          5700.5 406
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Las covariables `Center`, `Gender`, `DBP`, `Age`, `pre.BMI` parece que sí que afectan en la ganancia de fuerza en el test de repetición máxima. Sin embargo, el p-valor del coeficiente que contrasta si la combinación de haplotipos es significativa es superior a 0.05. No se puede decir que el gen "resistin" explique la ganancia de fuerza en el test de repetición máxima. En la Figura 62 se pensaba que sí que había diferencias para la ganancia de fuerza en el test de repetición máxima, pero parece que estas diferencias son explicadas por las otras covariables y no por el gen "resistin".

Modelo multivariante con `ND23_DIFF` y `NDRM_DIFF`

Para acabar, se realizará un modelo multivariante con las dos variables respuesta (ganancia de fuerza isométrica y para el test de repetición máxima). Se quiere ver la importancia conjunta ya que, para la primera variable, el gen "resistin" ha salido significativo y para la segunda no. Se han tenido en cuenta las interacciones de primer orden pero no se incluirán en el modelo final ya que ninguna ha salido significativa. Se realiza, en primer lugar, el modelo multivariante con los 4 niveles más frecuentes de los diplotipos.

Solamente las covariables significativas conjuntamente se incluyen en el modelo final que se muestra a continuación: `Center`, `Term`, `Género` y `DBP` y la combinación de haplotipos.

```
Type III MANOVA Tests: Pillai test statistic
                    Df test stat approx F num Df den Df    Pr(>F)
(Intercept)          1  0.044735  10.4196      2    445 3.781e-05 ***
Center               7  0.241813   8.7630     14    892 < 2.2e-16 ***
Term                 5  0.070821   3.2746     10    892 0.0003565 ***
Gender               1  0.050507  11.8355      2    445 9.817e-06 ***
DBP                  1  0.023759   5.4150      2    445 0.0047475 **
as.factor(orvhapl2)  4  0.019706   1.1095      8    892 0.3540593
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Las mismas covariables que en los otros modelos salen significativas también para el conjunto de las dos variables respuesta. No sorprende que la variable con los 4 diplotipos más frecuentes no salga significativa ya que cuando se ha contrastado para las dos variables por separado tampoco era significativa. Se contrastará el modelo con la variable con todos los diplotipos como predictora y las otras covariables.

```
Type III MANOVA Tests: Pillai test statistic
                   Df test stat approx F num Df den Df    Pr(>F)
(Intercept)         1  0.006529   1.2815      2    390  0.278796
Center              7  0.241630   7.6757     14    782 2.496e-15 ***
Term                5  0.059312   2.3900     10    782  0.008551 **
```

```
Gender              1  0.073467  15.4620      2    390 3.450e-07 ***
DBP                 1  0.027302   5.4734      2    390  0.004526 **
as.factor(orvhapl) 59  0.295301   1.1480    118    782  0.149493
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La combinación de haplotipos no es significativa, ya que el p-valor asociado a su coeficiente es superior a 0.05. No se puede afirmar que el gen "resistin" contribuya a la ganancia de fuerza de las dos variables conjuntamente.

Conclusiones

A modo de conclusión, se había observado que cuatro de los seis polimorfismos que se tienen del gen "resistin" contribuían a la ganancia de fuerza. Se ha estudiado si la combinación de los haplotipos del mismo son significativos por separado para las dos variables y después conjuntamente. Se ha visto este gen sí que influye en la ganancia de fuerza isómetrica pero no en la ganancia de fuerza en el test de repetición máxima. Se tienen evidencias para creer que el haplotipo que más influye en la ganancia de fuerza isométrica es el "CTACGA", podrían realizarse más pruebas que lo confirmaran. Este gen no es significativo para la explicación de las dos variables conjuntamente. Cabe remarcar, que todos los modelos han sido correctamente validados. Se recuerda que estas conclusiones son solamente para las personas de raza caucásica, debería estudiarse cómo se comporta el resto.

X. Conclusiones

El objetivo principal de este estudio era averiguar los factores genéticos que determinan la fuerza, el tamaño muscular y la respuesta al ejercicio. Se ha llevado a cabo mediante modelos lineales univariantes y multivariantes con covariables y SNPs y haplotipos del gen "resistin". Se pretendía también llegar a ser capaces de predecir la respuesta de los individuos al ejercicio e identificar qué características tienen los sujetos más susceptibles a ganar o perder fuerza. Este estudio podría ayudar a desarrollar agentes farmacológicos capaces de aumentar el tamaño y fuerza muscular con diversas aplicaciones, especialmente en el ámbito deportivo. También se quería averiguar qué otras covariables afectan a la ganancia de fuerza y al tamaño muscular.

Se ha estudiado la fuerza dinámica e isométrica de los flexores de los antebrazos antes y después de un entrenamiento de 12 semanas en el brazo no dominante. Finalmente se han podido estudiar 632 individuos de los 1396 que se tenían y 304 variables de las 346 de la base de datos original. Como características principales de los individuos se puede destacar que el estudio ha sido llevado a cabo en estudiantes (de entre 18 y 40 años) y con las restricciones explicadas en capítulo destinado al diseño del estudio. Se tiene una proporción de mujeres un poco más elevada que de hombres, se concentran entre los 20-25 y 22-27 años respectivamente. El 78% de los individuos estudiados son de raza caucásica; las razas africana, asiática e hispánica son minoritarias. No parece haber muchas diferencias entre el número de individuos estudiados entre los distintos centros. Contrariamente, en los segundos trimestres de cada año parece que se han investigado menos casos. Por último, se tiene una propoción de diestros muy superior a la de zurdos.

Se ha visto con las pruebas T-Student y F-Fisher y comparando el brazo no dominante con el dominante, que ha servido de control, que el entrenamiento llevado a cabo ha sido efectivo y que la mayoría de individuos han ganado fuerza en los distintos test. También se ha podido observar que han disminuido la sección transversal de bíceps y tríceps. No se ha podido demostrar que los individuos que ganan fuerza en el test de repetición máxima también lo hacen en el de fuerza isométrica.

La riqueza de los polimorfismos de la base de datos del estudio FAMuSS es buena, ya que la mayoría de los SNPs tienen una MAF alta y, por lo tanto, son diversos, cosa que ha permitido un buen estudio de los datos gracias a su variabilidad. Los 11 SNPs monomórficos que se tenían ya se habían eliminado del estudio con anterioridad. Para mirar la calidad de los polimorfismos se ha utilizado la Ley de Hardy-Weinberg, en las razas caucásica y asiática parecía que no todos los SNPs cumplían el equilibro. Se ha exluido del estudio el único polimorfismo que tenía un p-valor asociado en la raza caucásica por debajo del valor crítico que marca la Corrección de Bonferroni. El nivel de significación de las pruebas del estudio es de 0.05 aunque también se ha tenido en cuenta, como se ha dicho, la Corrección de Bonferroni y el *False Discovery Rate* cuando se tenía un número elevado de test. El análisis de componentes principales de los datos genéticos ha servido para ver que la población no es homogénea, se han podido observar distintas nubes de puntos según las razas en la primera, segunda y tercera dimensión.

Como se ha visto, se han realizado muchos esfuerzos para la estandarización entre centros y trimestres, como por ejemplo, se ha utilizado el mismo protocolo en los distintos lugares y momentos del tiempo, el material ha sido comprado a los mismos fabricantes, se han usado las mismas técnicas de medición, se ha establecido un laboratorio central en los que se examinan las muestras biológicas y que analizan las imágenes de resonancia magnética, entre otras medidas. A pesar de todo esto, los factores que hacen aumentar la variabilidad no han podido controlarse. Se han observado diferencias entre centros y entre trimestres a lo largo de todo el estudio. Para solventar este problema, se han incluido ambas variables en los modelos.

En el capítulo del diseño del estudio se ha explicado que la ingesta dietética no se ha controlado durante el estudio ya que se ha considerado que sería similar entre todos los individuos. No se han incluido sujetos con dietas raras y se han eliminado los individuos que han perdido o ganado 12 kg durante el estudio. Se desconoce si haber controlado la dieta hubiera hecho variar los resultados ya que las medidas tomadas para la estandarización de este factor no son muy estrictas.

Solamente se han podido analizar 632 individuos de los cuales 496 son de raza caucásica. Se ha visto en distintas pruebas que las poblaciones no son homogéneas. Aunque se han considerado distintas razas, la mayoría están poco representadas. Además, los individuos estudiados son voluntarios que viven en Estados

Unidos y pueden no representar a la población general ya que, por ejemplo, el criterio de participar en un entrenamiento diario durante 12 semanas puede afectar a la población del estudio. Se han tenido en cuenta 193 SNPs de los 225 candidatos escogidos por el estudio FAMuSS de los más de 10 millones descubiertos en el genoma humano y de los 20 millones que se ha estimado que existen (Ramírez-Bello, et al., 2012). También se han considerado las 17 covariables que se han creído más importantes para ayudar a explicar la ganancia de fuerza muscular de las muchas que existen, 5 covariables no han podido incluirse en el modelo ya que no se tenía información clara disponible de su significado. Se han tenido en cuenta las interacciones de primer orden entre covariables, que en ningún caso han resultado significativas. No obstante, se desconoce si pudiera existir cualquier otro tipo de interacción mucho más compleja. Como en cualquier estudio estadístico, hay tantos factores que pueden influir en una variable respuesta que no todos pueden tenerse en cuenta. Además, este estudio se ha visto limitado por las variables e individuos que han se registrado en la base de datos del estudio FAMuSS.

Uno de los principales problemas que se ha tenido en la depuración de los datos y en el estudio en general ha sido la gran cantidad de datos faltantes que tenía la base de datos original, casi un 40%. Se ha intentado solventar, como se ha explicado, con la eliminación de los cinco últimos trimestres, que se tenían alrededor de un 60% de *missings* y de las variables e individuos con más de un 50% de datos faltantes. No se ha podido ver a qué se debe que haya tantos faltantes. Sí que es cierto que con la depuración que se ha llevado a cabo se ha perdido algo de información pero se cree que es peor la imputación de *missings* no aleatorios que debería hacerse para algunas partes del estudio que la eliminación de estas variables e individuos. La base de datos estudiada tiene 623 individuos y 304 variables y un 8% de *missings*. Este pequeño porcentaje de datos faltantes que quedaban sí que se ha considerado que eran *Missing Completely Random* y se han imputado para las variables genéticas según una distribución trinomial donde era necesario. Para poder hacer un análisis de componentes principales, se necesita una matriz de datos completos. Sin embargo, en los modelos de regresión los registros con un dato faltante se descartan automáticamente.

Se han realizado cuatro modelos distintos para las variables de rendimiento muscular más importantes: la ganancia de fuerza isométrica y la ganancia de fuerza en el test de repetición máxima. Se han considerado los modelos aditivo, recesivo, dominante y codominante en ambas variables respuesta. El número de polimorfismos asociados a estas variables es bastante grande. En la Tabla 6 se han mostrado los polimorfismos significativos de los cuatro modelos, son los SNPs que más influyen en la ganancia de fuerza isométrica. El SNP más influyente en este test es "vdr_tq1" y el único con un p-valor por debajo del valor de la Corrección de Bonferroni. Otros tres polimorfismos de este gen, "vdr", también son significativos. También se encuentran otros genes que parecen tener más de un SNP asociado a la ganancia de fuerza isométrica son "akt1", "resistin" y "cast". En cuanto a los polimorfismos significativos en la ganancia de fuerza del test de fuerza isométrica, se listan en la Tabla 11. El SNP más asociado a la ganancia de fuerza en este test es el "b2b". El único gen que tiene como significativo más de un polimorfismo es el gen "resistin". Estos genes identificados son distintos a los que se sabía que eran importantes en los animales de granja (Foulkes, 2009).

Se ha observado que el gen "resistin" aparecía como significativo en cuatro genes distintos entre las dos variables respuesta. Para los individuos de la raza caucásica, se ha elaborado un modelo con los cuatro diplotipos más comunes como variable respuesta junto con las otras covariables y otro con todos los diplotipos, incluyendo los minoritarios, que se tenían para las dos variables respuesta y para el modelo multivariante. Mediante la estimación de haplotipos de este gen se ha visto que sí que afecta en la ganancia de fuerza isómetrica pero no se ha podido demostrar asociación en la ganancia de fuerza en el test de repetición máxima ni para el modelo multivariante con las dos variables respuestas conjuntamente. Se ha visto que el haplotipo "CTACGA" está asociado a la ganancia de fuerza isométrica. Se ha buscado información sobre polimorfismos y genes más importantes y, como este estudio no se había llevado a cabo antes, no se ha encontrado ningún estudio que los asocie con la ganancia de fuerza.

Aunque no era el objetivo principal del estudio también se ha visto qué covariables son las que globalmente son más importantes para la ganancia de fuerza isométrica y en el test de repetición máxima. Se ha visto que los hombres, las personas con presión sanguínea diastólica alta, las que tienen índice de masa corporal alto y las que tienen menos edad son las que tienden a tener más ganancia de fuerza. Los individuos que tienen más dificultades al ganar fuerza en media son los de raza asiática y los que tienen una cantidad de lipoproteínas de muy baja densidad alta. Se han encontrado diferencias entre distintos centros y trimestres,

las personas estudiadas en *University of West Virginia* y en *University of Massachusetts* tienden a ganar más fuerza, mientras que las estudiadas en el trimestre "03-1" menos. Todos estos factores también han de tenerse en cuenta a la hora de predecir la ganancia de fuerza de un individuo.

De cara a posibles análisis futuros, podría obtenerse una muestra más grande de las distintas etnias, ya que la única que ha podido estudiarse a fondo ha sido la caucásica. Por otro lado, también deberían estudiarse las variables que han tenido que descartarse por la gran cantidad de *missings* o añadir algunas nuevas como si el individuo es fumador o si ha practicado deporte de alto rendimiento a lo largo de su vida, que puede afectar a la entrenabilidad del sujeto. También podrían tomarse las variables de rendimiento muscular y realizar los test en algún punto medio del estudio para ver así cómo evolucionan. Solamente se han estudiado dos de las 76 variables respuesta que se podían tener, además algunas han tenido que descartarse a causa de la falta de información que se tenía. Podría tratarse de estudiar algunas otras y comparar los resultados. Por otro lado, a nivel de modelización, se han ajustado inicialmente los SNPs uno por uno. Podría probarse algo más complejo como poner como predictor más de un polimorfismo a la vez o incluso con sus interacciones. Se han estimado haplotipos del gen "resistin", que se creía que era el más importante, sin embargo han quedado otros como "akt1", "vdr" y "cast" que también se tiene indicios de que podrían ser influyentes. La parte de regresión multivariante también podría extenderse mucho más. Quizás primero se deberían estudiar las relaciones entre respuestas para luego poder aplicar métodos de regresión multivariante, manova, correlaciones canónicas o biplots entre otros.

XI. Glosario

ADN "Ácido desoxirribonucleico. Biopolímero cuyas unidades son desoxirribonucleótidos y que constituye el material genético de las células y contiene en su secuencia la información para la síntesis de proteínas, conocido más por sus siglas ADN o DNA". Real Academia Española. *Diccionario de la lengua española* Madrid: Espasa Libros, 2014. ISBN: 9788467047882

AIC "El criterio de información de Akaike (AIC) es una medida de la calidad relativa de un modelo estadístico, para un conjunto dado de datos. Como tal, el AIC proporciona un medio para la selección del modelo.

$$AIC = 2k - 2ln(L)$$

donde k es el número de parámetros en el modelo estadístico y L es el máximo valor de la función de verosimilitud para el modelo estimado. El mejor modelo es el que minimiza el AIC". Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Criterio_de_información_de_Akaike

Alelo "Cada una de las formas alternativas de un gen que ocupan el mismo lugar en los cromosomas homólogos y cuya expresión determina las características del mismo rasgo de organización, como el color de los ojos." Real Academia Española. *Diccionario de la lengua española* Madrid: Espasa Libros, 2014. ISBN: 9788467047882

Alometria "En biología la alometría se refiere a los cambios de dimensión relativa de las partes corporales correlacionados con los cambios en el tamaño total. Más específicamente durante el desarrollo de un organismo, la alometría en el crecimiento, se refiere al crecimiento diferencial de diferentes partes del cuerpo." Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Alometría

Análisis de componentes principales (ACP) "El ACP es una técnica utilizada para describir un conjunto de datos en términos de nuevas variables ("componentes") no correlacionadas. Los componentes se ordenan por la cantidad de varianza original que describen, por lo que la técnica es útil para reducir la dimensionalidad de un conjunto de datos. Técnicamente, el ACP busca la proyección según la cual los datos queden mejor representados en términos de mínimos cuadrados. Esta convierte un conjunto de observaciones de variables posiblemente correlacionadas en un conjunto de valores de variables sin correlación lineal llamadas componentes principales". Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Análisis_de_componentes_principales

Andrógenos "Hormonas sexuales masculinas que corresponden a la testosterona, la androsterona y la androstenediona" Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Andrógeno

ANOVA "El análisis de la varianza, es una colección de modelos estadísticos y sus procedimientos asociados, en el cual la varianza está particionada en ciertos componentes debidos a diferentes variables explicativas." Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Análisis_de_la_varianza

Antebrazo "Parte del brazo desde el codo hasta la muñeca." Real Academia Española. *Diccionario de la lengua española* Madrid: Espasa Libros, 2014. ISBN: 9788467047882

Antropométrico "Perteneciente o relativo al estudio de las proporciones y medidas del cuerpo humano." Real Academia Española. *Diccionario de la lengua española* Madrid: Espasa Libros, 2014. ISBN: 9788467047882

Área transversal "Área de aquellos planos que son perpendiculares al eje longitudinal de una estructura". Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Plano_transverso

ARN "Biopolímero cuyas unidades son ribonucleótidos y que, según su función, puede ser mensajero, ribosómico o de transferencia." Real Academia Española. *Diccionario de la lengua española* Madrid: Espasa Libros, 2014. ISBN: 9788467047882

ARNm "Molécula de ácido ribonucleico que transmite la información genética del ADN para ser traducida durante la síntesis de proteínas." Real Academia Española. *Diccionario de la lengua española* Madrid: Espasa Libros, 2014. ISBN: 9788467047882

Axial "Perteneciente o relativo al eje." Real Academia Española. *Diccionario de la lengua española* Madrid: Espasa Libros, 2014. ISBN: 9788467047882

Bíceps "Músculo bíceps que va desde el omóplato a la parte superior del radio y que, al contraerse, dobla el antebrazo sobre el brazo." Real Academia Española. *Diccionario de la lengua española* Madrid: Espasa Libros, 2014. ISBN: 9788467047882

Brazo dominante Brazo con el que un individuo se siente más cómodo al realizar tareas que requieran habilidad.

Brazo no dominante Brazo con el que un individuo se siente menos cómodo al realizar tareas que requieran habilidad.

Coeficiente de determinación (R^2) "El coeficiente de determinación, se define como la proporción de la varianza total de la variable explicada por la regresión. El coeficiente de determinación, también llamado R cuadrado, refleja la bondad del ajuste de un modelo a la variable que pretender explicar. Se calcula como

$$R^2 = \frac{\sum_{t=1}^{T}(\hat{Y}_t - \bar{Y})^2}{\sum_{t=1}^{T}(Y_t - \bar{Y})^2}$$

". Economipedia [consulta a 31 de mayo de 2019] Disponible en: https://economipedia.com/definiciones/r-cuadrado-coeficiente-determinacion.html

Coeficiente de determinación ajustado (R^2_{adj}) "El coeficiente de determinación ajustado es la medida que define el porcentaje explicado por la varianza de la regresión en relación con la varianza de la variable explicada. Es decir, lo mismo que el R cuadrado, pero con una diferencia. Esa diferencia se encuentra en que el coeficiente de determinación ajustado penaliza la inclusión de variables. Se calcula como

$$\bar{R}^2 = 1 - \frac{N-1}{N-k-1}[1 - R^2]$$

donde N es el tamaño de la muestra y k el número de variables explicativas. Economipedia [consulta a 31 de mayo de 2019] Disponible en: https://economipedia.com/definiciones/r-cuadrado-coeficiente-determinacion.html

Corrección de Bonferroni "En estadística, la Corrección de Bonferroni es uno de los varios métodos utilizados para contrarrestar el problema de las comparaciones múltiples. (...) La prueba de hipótesis estadística se basa en rechazar la hipótesis nula si la probabilidad de los datos observados en las hipótesis nulas es baja. Si se prueban múltiples hipótesis, aumenta la probabilidad de un evento raro y, por lo tanto, aumenta la probabilidad de rechazar incorrectamente una hipótesis nula (es decir, cometer un error de tipo I.). La Corrección de Bonferroni compensa ese aumento al probar cada hipótesis individual en un nivel significativo de $\frac{\alpha}{m}$, donde α es el nivel de alfa general deseado y m es el número de hipótesis". Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Corrección_de_Bonferroni

Covariable "En estadística, una covariable es una variable que posiblemente predice el resultado bajo estudio. Una covariable puede ser de interés directo o puede ser una variable de confusión o con interacción." Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Covariable

Creatina "La creatina (también denominada α-metil guanido-acético y frecuentemente abreviado en la literatura como Cr) es un ácido orgánico nitrogenado que se encuentra en los músculos y células nerviosas de algunos organismos vivos. Es un derivado de los aminoácidos (molécula orgánica) muy parecido a ellos en cuanto a su estructura molecular". Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Creatina

Cromosoma "Filamento condensado de ácido desoxirribonucleico, visible en el núcleo de las células durante la mitosis y cuyo número es constante para las células de cada especie animal o vegetal." Real Academia Española. *Diccionario de la lengua española* Madrid: Espasa Libros, 2014. ISBN: 9788467047882

Diplotipo Un diplotipo consiste la combinación de dos alelos, uno materno y otro paterno.

Esquema aditivo Esquema en el que los componentes, en este caso los tres niveles de cada polimorfismo, se agregan de manera conjunta para modelar los datos. Genética médica blog [consulta 19 de junio 2019] Disponible en: https://revistageneticamedica.com/blog/grupos-sanguineos/

Esquema codominante Esquema en el cual ambos alelos se expresan simultáneamente en el heterocigoto y dominan por igual.

Esquema dominante y esquema recesivo En genética, la dominancia es una relación entre alelos de un mismo gen, en el que uno enmascara la expresión, siendo posible tres combinaciones de alelos -genotipo AA, Aa y aa. Si los individuos homocigóticos AA y aa muestran diferentes formas para una característica y los individuos heterocigóticos Aa son idénticos al fenotipo de los individuos AA, entonces el alelo A se dice que domina, que es dominante o que muestra dominancia sobre el alelo a, y a se dice que es recesivo con respecto a A. Es un concepto clave en las leyes de Mendel y en la genética clásica. Muchas veces el alelo dominante fabrica códigos por una proteína funcional mientras el alelo recesivo no lo hace. Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Dominancia_(genética)

Estandarización "La normalización de índices significa ajustar los valores medidos en diferentes escalas respecto a una escala común, a menudo previo a un proceso de realizar promedios". Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Normalización_(estadística)

Exón "El exón es la región de un gen que no es separada durante el proceso de corte y empalme y, por tanto, se mantienen en el ARN mensajero maduro. En los genes que codifican una proteína, son los exones los que contienen la información para producir la proteína codificada en el gen". Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Exón

Factor Un factor es una variable categórica con un número finito de valores o niveles.

False Discovery Rate "Método para estimar la proporción esperada de falsos positivos de entre los test considerados como significativos. El objetivo de controlar el *False Discovery Rate* es establecer un límite de significancia para un conjunto de test tal que, de entre todos los test considerados como significativos, la proporción de hipótesis nulas verdaderas (falsos positivos) no supere un determinado valor". Rpubs [consulta 20 de junio de 2019]. Disponible en: https://rpubs.com/Joaquin_AR/236898

Falso positivo "En un estudio de investigación experimental, al error de tipo I, también llamado erróneamente error de tipo alfa (α). Es el error que se comete cuando el investigador rechaza la hipótesis nula (H_0: el supuesto inicial) siendo esta verdadera en la población". Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Errores_de_tipo_I_y_de_tipo_II

Fenotípico "Perteneciente o relativo a la manifestación variable del genotipo de un organismo en un determinado ambiente". Real Academia Española. *Diccionario de la lengua española* Madrid: Espasa Libros, 2014. ISBN: 9788467047882

Flebotomía "Se llama flebotomía a varios procedimientos relacionados con la sangre, pero por lo general este término se atribuye a una modalidad de tratamiento médico que consiste en la extracción de sangre del paciente para el tratamiento de dolencias". Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Sangría_(tratamiento_médico)

Fuerza dinámica Fuerza en la que se crea un desplazamiento de los músculos en el que se tiene que observar claramente un movimiento externo al realizar cualquier acto.

Fuerza excéntrica Fuerza en la que el músculo continúa contraído y las inserciones musculares se distancian, el movimiento se genera es a favor de la gravedad. La contracción excéntrica tiene la importante función de controlar, de "frenar" el movimiento cuando va a favor de la gravedad.

Fuerza isométrica "La fuerza isométrica hace referencia a la tensión de un músculo y su mantenimiento en una posición estacionaria al tiempo que se mantiene la tensión. (...) Puede llamarse también ejercicio isométrico a aquel en el cual se aplica una fuerza a un objeto que opone resistencia". Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Ejercicio_isométrico

Gen "Secuencia de ADN que constituye la unidad funcional para la transmisión de los caracteres hereditarios". Real Academia Española. *Diccionario de la lengua española* Madrid: Espasa Libros, 2014. ISBN: 9788467047882

Genotipado "Se entiende el proceso de determinación del genotipo de una variante en el ADN específico de un organismo biológico, mediante un procedimiento de laboratorio". Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Genotipificar

Haplotipo "Un haplotipo en genética es una combinación de alelos de diferentes *locus* de un cromosoma que son transmitidos juntos. Un haplotipo puede ser un locus, varios *loci*, o un cromosoma entero dependiendo del número de eventos de recombinación que han ocurrido entre un conjunto dado de *loci.*" Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Haplotipo

Homocedasticidad "En estadística se dice que un modelo predictivo presenta homocedasticidad cuando la varianza del error condicional a las variables explicativas es constante a lo largo de las observaciones". Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Homocedasticidad)

Húmero "Hueso del brazo, que se articula por uno de sus extremos con la escápula y por el otro con el cubito y el radio". Real Academia Española. *Diccionario de la lengua española* Madrid: Espasa Libros, 2014. ISBN: 9788467047882

Imagen por resonancia magnética "Una imagen por resonancia magnética (IRM) es una técnica no invasiva que utiliza el fenómeno de la resonancia magnética nuclear para obtener información sobre la estructura y composición del cuerpo a analizar. Esta información es procesada por ordenadores y transformada en imágenes del interior de lo que se ha analizado". Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Imagen_por_resonancia_magnética)

Imputación de datos faltantes Sustitución de los valores *missing* por valores posibles en la base de datos de un estudio estadístico.

Intrón "Un intrón es una región del ADN que forma parte de la transcripción primaria de ARN, pero a diferencia de los exones, son eliminados del transcrito maduro, previamente a su traducción." Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Intrón

Lipoproteína "Proteína conjugada cuyos componentes no proteínicos son lípidos". Real Academia Española. *Diccionario de la lengua española* Madrid: Espasa Libros, 2014. ISBN: 9788467047882

Locus "Un locus (en plural *loci*) es una posición fija en un cromosoma, que determina la posición de un gen o de un marcador genético. En biología, y, por extensión, en computación evolutiva, se le usa para identificar posiciones de interés sobre determinadas secuencias." Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Locus

Media "En matemáticas y estadística, la media aritmética, también llamada promedio o media, de un conjunto infinito de números es el valor característico de una serie de datos cuantitativos, objeto de estudio que parte del principio de la esperanza matemática o valor esperado, se obtiene a partir de la suma de todos sus valores dividida entre el número de sumandos". Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Media_aritmética

Mediana "En el ámbito de la estadística, la mediana representa el valor de la variable de posición central en un conjunto de datos ordenados." Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Mediana_(estadística)

Medida normalizada "Se refiere a la creación de versiones estadísticas cambiadas y escaladas, donde la intención es que los valores normalizados permitan la comparación de los valores normalizados con

conjuntos de datos de manera que elimine los efectos de influencias". Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Normalización_(estadística)

Missing "Son aquellos que no constan debido a cualquier acontecimiento, como por ejemplo errores en la transcripción de los datos o la ausencia de disposición a responder a ciertas cuestiones de una encuesta." UV [consulta a 31 de mayo de 2019] Disponible en: https://www.uv.es/webgid/Descriptiva/23_valores_faltantes.html

Monomórfico (SNP) Un SNP monomórfico hace referencia a la existencia en una sola forma alélica de un gen.

Nivel basal Nivel del que se parte, de base.

Outlier Valor atípico, "es una observación que es numéricamente distante del resto de los datos." Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Valor_atípico)

Polimorfismo "Propiedad de las especies de seres vivos cuyos individuos pueden presentar diferentes formas o aspectos, bien por diferenciarse en castas, como las termitas, bien por tratarse de distintas etapas del ciclo vital, como la oruga y la mariposa". Real Academia Española. *Diccionario de la lengua española* Madrid: Espasa Libros, 2014. ISBN: 9788467047882

Presión sanguínea diastólica "Es la tensión ejercida por la sangre durante la fase diastólica que circula sobre las paredes de los vasos sanguíneos, y constituye uno de los principales signos vitales.(...) Es el valor mínimo de la curva de presión". Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Presión_sanguínea

Presión sanguínea sistólica "Es la tensión ejercida por la sangre durante la fase sistólica que circula sobre las paredes de los vasos sanguíneos, y constituye uno de los principales signos vitales. (...) El máximo de la curva de presión en las arterias y que ocurre cerca del principio del ciclo cardíaco". Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Presión_sanguínea

Proteína "Sustancia constitutiva de la materia viva, formada por una o varias cadenas de aminoácidos". Real Academia Española. *Diccionario de la lengua española* Madrid: Espasa Libros, 2014. ISBN: 9788467047882

P-valor "En contrastes de hipótesis y en estadística general, el valor de p se define como la probabilidad correspondiente al estadístico de ser posible bajo la hipótesis nula. Si cumple con la condición de ser menor al nivel de significancia impuesto arbitrariamente, entonces la hipótesis nula será, eventualmente, rechazada." Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Valor_p

Q-Q Plot "En estadística, un gráfico Q-Q ("Q" viene de cuantil) es un método gráfico para el diagnóstico de diferencias entre la distribución de probabilidad de una población de la que se ha extraído una muestra aleatoria y una distribución usada para la comparación. Una forma básica de gráfico surge cuando la distribución para la comparación es una distribución teórica". Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Gráfico_Q-Q

Significación estadística "En estadística, un resultado o efecto es estadísticamente significativo cuando es improbable que haya sido debido al azar. Una"diferencia estadísticamente significativa" solamente significa que hay evidencias estadísticas de que hay una diferencia; no significa que la diferencia sea grande, importante o radicalmente diferente. El nivel de significación de una prueba estadística es un concepto estadístico asociado a la verificación de una hipótesis. En pocas palabras, se define como la probabilidad de tomar la decisión de rechazar la hipótesis nula cuando ésta es verdadera (decisión conocida como error de tipo I, o "falso positivo"). La decisión se toma a menudo utilizando el valor p: si el valor p es inferior al nivel de significación, entonces la hipótesis nula es rechazada. Cuanto menor sea el valor p, más significativo será el resultado" Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Significación_estadística

SNP "Un polimorfismo de un solo nucleótido o SNP (*Single Nucleotide Polymorphism, pronunciado snip*) es una variación en la secuencia de ADN que afecta a una sola base (adenina (A), timina (T), citosina (C)

o guanina (G)) de una secuencia del genoma". Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Polimorfismo_de_nucleótido_único

Supina "Posición de una persona tendida sobre el dorso, o de la mano con la palma hacia arriba". Real Academia Española. *Diccionario de la lengua española* Madrid: Espasa Libros, 2014. ISBN: 9788467047882

Tríceps "Músculo tríceps unido al fémur y la tibia y que al contraerse extiende con fuerza la pierna". Real Academia Española. *Diccionario de la lengua española* Madrid: Espasa Libros, 2014. ISBN: 9788467047882

Valor atípico *Outlier*, "es una observación que es numéricamente distante del resto de los datos". Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Valor_atípico.

XII. Bibliografía

12.1 Fuente de los datos

[1] Reilly, C. S. [consulta a 14 de marzo de 2019] Disponible en: http://www.biostat.umn.edu/~cavanr/FMS_data.txt

12.2 Fuentes escritas

[2] American College of Sports Medicine position stand (ACSM). *Progression models in resistance training for healthy adults*. Nueva York: Medicine and Science in Sports and Exercise 41, 687-708, 2009. DOI: 10.1249/MSS.0b013e3181915670.

[3] Beunen, G. y Thomis, M. *Gene powered? Where to go from heritability (h2) in muscle strength and power?*. Leuven: Exercise and Sport Science Reviews 32, 148-154, 2009. PMID: 15604933

[4] Cascales Salinas, B.; Lucas Saorín, P.; Mira Ros, J.M.; Pallarés Ruiz, A.J.; Sánchez-Pedreño Guillén, S. *El libro de LATEX*. Murcia: Prentice Hall, 2003. ISBN: 84-205-3779-9.

[5] Foulkes, Andrea S. *Applied Statistical Genetics with R for Population-based Association Studies*. Nueva York: Springer Science + Business Media, 2009. ISBN: 978-0-387-89553-6.

[6] Graffelman, J. *Haplotype Estimation (Phasing)*. Barcelona, 2018.

[7] Ramírez Bello, J.; Vargas-Alarcón, G.; Tovilla-Zárate, C. y Fragoso, J.M. *Polimorfismos de un solo nucleótido (SNP): implicaciones funcionales de los SNP reguladores (rSNP) y de los SNP-ARN estructurales (srSNP) en enfermedades complejas*. México: Gaceta Médica de México 149, 220-228, 2013. ISSN: 0016-3813

[8] Real Academia Española. *Diccionario de la lengua española* Madrid: Espasa Libros, 2014. ISBN: 9788467047882

[9] Sinnwell, Jason P. y Schaid, J.D. *Haplo Stats, Statistical Methods for Haplotypes when Linkage Phase is Ambiguous*. Mayo Clinic Division of Health Sciences Research, 2016.

[10] Stewart, C.E. y Rittweger, J. *Adaptive processes in skeletal muscle: molecular regulators and genetic influences*. Cologne: Journal of Musculoskeletal and Neuronal Interactions 66, 73-86, 2006. PMID:16675891

[11] Thomson, P.D. Functional Polymorphisms Associated with Human Muscle Size and Strength. A: *Medicine & Science in Sports & Exercise*, Vol. 36, No. 7, p.1132-1139, 2004. ISSN: 0195-9131

[12] Universitat de Barcelona, Universitat Politècnica de Catalunya. *Instruccions per a l'eleaboració i difusió del treball final de grau*

12.3 Fuentes multimedia

[13] Centre de Recursos per a l'Aprenentatge i la Investigació [consulta 8 de junio de 2019]. Disponible en: http://crai.ub.edu/ca/que-ofereix-el-crai/citacions-bibliografiques/documents-electronics

[14] Economipedia [consulta a 31 de mayo de 2019] Disponible en: https://economipedia.com/definiciones/r-cuadrado-coeficiente-determinacion.html

[15] Genética médica blog [consulta 19 de junio 2019] Disponible en: https://revistageneticamedica.com/blog/grupos-sanguineos/

[16] Rpubs [consulta 8 de junio de 2019]. Disponible en: https://rpubs.com/gabrielmartos/multivPCA

[17] Rpubs [consulta 20 de junio de 2019]. Disponible en: https://rpubs.com/Joaquin_AR/236898

[18] Rpubs [consulta 8 de junio de 2019]. Disponible en: https://rpubs.com/Joaquin_AR/287787

[19] UV [consulta a 31 de mayo de 2019] Disponible en: https://www.uv.es/webgid/Descriptiva/23_valores_faltantes.html

[20] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Alometría

[21] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Análisis_de_componentes_principales

[22] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Análisis_de_la_varianza

[23] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Andrógeno

[24] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Corrección_de_Bonferroni

[25] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Covariable

[26] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Criterio_de_información_de_Akaike

[27] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Creatina

[28] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Dominancia_(genética)

[29] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Ejercicio_isométrico

[30] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Errores_de_tipo_I_y_de_tipo_II

[31] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Exón

[32] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Genotipificar

[33] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Gráfico_Q-Q

[34] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Haplotipo

[35] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Homocedasticidad)

[36] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Imagen_por_resonancia_magnética)

[37] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Intrón

[38] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Locus

[39] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Media_aritmética

[40] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Mediana_(estadística)

[41] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Normalización_(estadística)

[42] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Plano_transverso

[43] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Polimorfismo_de_nucleótido_único

[44] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Presión_sanguínea

[45] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Sangría_(tratamiento_médico)

[46] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Significación_estadística

[47] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Valor_atípico)

[48] Wikipedia [consulta a 31 de mayo de 2019] Disponible en: https://es.wikipedia.org/wiki/Valor_p

12.4 Paquetes de R

[49] Stoffer, D. astsa (versión 1.9) https://www.rdocumentation.org/packages/astsa/versions/1.9/topics/FDR [consulta a 20 de junio de 2019]

[50] base (versión 3.4.3) https://stat.ethz.ch/R-manual/R-devel/library/base/html/00Index.html [consulta a 10 de mayo de 2019]

[51] gap (versión 1.1-22) https://cran.r-project.org/web/packages/gap/index.html [consulta a 23 de abril de 2019]

[52] haplo.stats (versión 1.7.9) https://cran.r-project.org/web/packages/haplo.stats/index.html [consulta a 10 de junio de 2019]

[53] Graffelman, J. Hardy Weinberg (versión 1.6.1) https://cran.r-project.org/web/packages/HardyWeinberg/index.html [consulta a 11 de abril de 2019]

[54] Sarkar, D. lattice (versión 0.28-38) https://cran.r-project.org/web/packages/lattice/index.html [consulta 20 de mayo de 2019]

[55] MASS (versión 7.3-48) https://cran.r-project.org/web/packages/MASS/index.html [consulta a 28 de abril de 2019]

[56] stats (versión 3.4.3) https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html [consulta a 27 de abril de 2019]

XIII. Lista de siglas

1RM Una Repetición Máxima

ACP Análisis de Componentes Principales

ADN Ácido Desoxirribonucleico

AIC *Akaike Information Criterion*

ANOVA *Analysis Of Variance*

cc Centímetros cúbicos

CSA *Cross-Sectional Area*

FAMuSS *Functional Single Nucleotide Polymorfisms Associated with Muscle Size and Strength*

FDR *False Discovery Rate*

IRM Imagen por Resonancia Magnética

MAF *Minor Allele Frequency*

MCAR *Missing Completely Random*

NA *Not Available*

NEX Número de excitaciones

NIH *National Institutes of Health*

PCA *Principal Component Analysis*

POM *Point Of mesure*

ROI *Regions Of Interest*

SNP *Single Nucleotide Polumorfisms*

XIV. Lista de figuras

XIV. Lista de tablas

XV. Anexo

```
knitr::include_graphics('/home/blanca/Documentos/TFG/Adj/pic1_2.png')

knitr::include_graphics('/home/blanca/Documentos/TFG/Adj/pic2_1.png')

knitr::include_graphics('/home/blanca/Documentos/TFG/Adj/tabla1_1.png')

knitr::include_graphics('/home/blanca/Documentos/TFG/Adj/tabla2_1.png')


## V. SELECCION DE VARIABLES E INDIVIDUOS, CAMBIO DE UNIDADES Y RECATEGORIZACION

#rawdata <- read.delim("/home/blanca/Documentos/cole/4 ESTADISTICA/TFG/FMS_data.txt",
#                      header = TRUE, sep = "\t")

#rawdata <- read.delim("C:/Users/Blanca/Documents/TFG/FMS_data.txt",
#                     header = TRUE, sep = "\t")

rawdata <- read.delim("/home/blanca/Documentos/TFG/FMS_data.txt",
                        header = TRUE, sep = "\t")
#row.names(rawdata) <- rawdata$id
#has a duplicate id

which(rawdata$id == 'WV-1332')
rawdata <- rawdata[-1388,]

attach(rawdata)
row.names(rawdata) <- id
rawdata <- rawdata[, -1]

#View(rawdata)
#write.csv(rawdata, file = "/media/blanca/7111-CB28/4 ESTADISTICA/TFG/rawdata.csv")

rawdata[which(rawdata[, 'Race']=='Am␣Indian'), 'Race'] <- 'Other'

#rawdata[, CRP == '0.1'] <- as.factor('0.1 - 1')
#rawdata$CRP ['0.1' || '0.2' || '0.3' || '0.4' || '0.5' || '0.6' || '0.7' || '0.8' ||
  #'0.9']
#<- as.factor('0.1 - 1')

rawdata[, 'CRP'] <- car::recode(rawdata[, 'CRP'], "c(1,1.1,1.2,1.3,1.4,1.5,1.6,
1.7,1.8,18.6,2.5,5,6.5)␣=␣'>1';c(0.1,0.2,0.3,0.4,0,5,0.6,0.7,0.8,0.9)
='0.1-1';else='<0.1'")

rawdata <- rawdata[, -29]
detach(rawdata)

rawdata[, 'Pre.weight'] <- rawdata[, 'Pre.weight'] / 2.2046
rawdata[, 'Post.weight'] <- rawdata[, 'Post.weight'] / 2.2046

rawdata[, 'Pre.height'] <- rawdata[, 'Pre.height'] * 2.54
rawdata[, 'Post.Height'] <- rawdata[, 'Post.Height'] * 2.54

## FUNCION PARA BUSCAR MISSINGS
searchNA <- function(x){
  s <- 0
```

```
  for(i in 1:length(x)){
    if(is.na(x[i]==TRUE)){
      s <- s  + 1
    }
  }
  return(s*100/length(x))
}


## 5.1 PORCENTAJE DE DATOS FALTANTES POR VARIABLE
rawNAvar <- round(apply(rawdata, 2, searchNA),2)
#rawNAvar[1:10]

hist(rawNAvar, main='Histograma␣del
␣␣␣␣␣porcentaje␣de␣missings␣por␣variable',
     xlab='Porcentaje␣de␣missings', ylab='Numero␣de␣varibles', xlim = c(0, 100),
     ylim = c(0, 100), cex.main = 1, col = 'lightblue', las = 1, cex.lab = 0.8,
     cex.axis = 0.8)

rawNAvar[which(rawNAvar > 80)]
sum(rawNAvar > 80)


## 5.2 PORCENTAJE DE DATOS FALTANTES POR INDIVIDUO

rawNAind <- round(apply(rawdata, 1, searchNA),2)
#rawNAind[1:10]

hist(rawNAind, main='Histograma␣del
␣␣␣␣␣porcentaje␣de␣missings␣por␣individuo',
     xlab='Porcentaje␣de␣missings', ylab="Numero␣de␣individuos", xlim = c(0, 100),
     ylim = c(0, 420), cex.main = 1, col = 'lightblue', las = 1, cex.lab = 0.8,
     cex.axis = 0.8)

sum(rawNAind > 80)

## 5.3 PORCENTAJE DE DATOS FALTANTES TOTAL
mean(rawNAvar)

## 5.4 ELIMINACION DE VARIABLES E INDIVIDUOS CON MUCHOS MISSINGS
par(xpd=T, mar=par()$mar+c(0,0,0,6))
col <- rep('black', 1396)
col[which(rawdata[, 'Term']=='02-1')] <- 'gold'
col[which(rawdata[, 'Term']=='02-2')] <- 'chocolate1'
col[which(rawdata[, 'Term']=='02-3')] <- 'red'
col[which(rawdata[, 'Term']=='03-1')] <- 'brown4'
col[which(rawdata[, 'Term']=='03-2')] <- 'purple'
col[which(rawdata[, 'Term']=='03-3')] <- 'blue4'
col[which(rawdata[, 'Term']=='04-1')] <- 'cyan3'
col[which(rawdata[, 'Term']=='04-2')] <- 'gray50'
col[which(rawdata[, 'Term']=='04-3')] <- 'darkolivegreen3'
col[which(rawdata[, 'Term']=='05-1')] <- 'green4'
plot(1:1396, rawNAind, main = 'Grafico␣individuos␣vs␣porcentaje␣de␣missings',
     xlab = 'Individuos', ylab = 'Porcentaje␣de␣missings', pch = 19, col = col,
     cex = 0.8, ylim = c(0, 100), las = 1, cex.main = 1, cex.lab = 0.8, cex.axis = 0.8)
legend("topright", inset=c(-0.17,0), legend=c(c('02-1', '02-2', '02-3', '03-1', '03-2',
                                               '03-3', '04-1', '04-2', '04-3', '05-1',
                                               'NA')),
       pch=20, title="Term", xpd = TRUE, col = c('gold', 'chocolate1', 'red', 'brown4',
```

```
                                                    'purple', 'blue4', 'cyan3', 'gray50',
                                                    'darkolivegreen3', 'green4', 'black'),
                                                    cex = 0.9)
legend('bottomright', inset=c(-0.19), legend = 'Media', lty = 2, lwd =2, col = 'red',
cex = 0.7)
#abline(h = mean(NAind), col = 'red', lty = 2)
x <- c(-46:1452)
y <- c(rep(mean(rawNAind), 1499))
l <- list(x = x, y = y)
lines(l, lty = 2, col = 'red', lwd = 2)
par(mar=c(5, 4, 4, 2) + 0.1)


#NA en term
Xsub <- rawdata[is.na(rawdata$Term),1:224]
ndatos <- nrow(Xsub)*ncol(Xsub)
sum(is.na(Xsub))/ndatos

#NA total
Xgen <- rawdata[,1:224]
ndatos <- nrow(Xgen)*ncol(Xgen)
sum(is.na(Xgen))/ndatos

#NA por term
data1T <- Xgen[which(rawdata$Term=='02-1'), ]
NA1T <- sum(is.na(data1T))/(ncol(data1T)*nrow(data1T))
NA1T

data2T <- Xgen[which(rawdata$Term=='02-2'), ]
NA2T <- sum(is.na(data2T))/(ncol(data2T)*nrow(data2T))
NA2T

data3T <- Xgen[which(rawdata$Term=='02-3'), ]
NA3T <- sum(is.na(data3T))/(ncol(data3T)*nrow(data3T))
NA3T

data4T <- Xgen[which(rawdata$Term=='03-1'), ]
NA4T <- sum(is.na(data4T))/(ncol(data4T)*nrow(data4T))
NA4T

data5T <- Xgen[which(rawdata$Term=='03-2'), ]
NA5T <- sum(is.na(data5T))/(ncol(data5T)*nrow(data5T))
NA5T

data6T <- Xgen[which(rawdata$Term=='03-3'), ]
NA6T <- sum(is.na(data6T))/(ncol(data6T)*nrow(data6T))
NA6T

data7T <- Xgen[which(rawdata$Term=='04-1'), ]
NA7T <- sum(is.na(data7T))/(ncol(data7T)*nrow(data7T))
NA7T

data8T <- Xgen[which(rawdata$Term=='04-2'), ]
NA8T <- sum(is.na(data8T))/(ncol(data8T)*nrow(data8T))
NA8T

data9T <- Xgen[which(rawdata$Term=='04-3'), ]
NA9T <- sum(is.na(data9T))/(ncol(data5T)*nrow(data9T))
NA9T
```

```
data10T <- Xgen[which(rawdata$Term=='05-1'), ]
NA10T <- sum(is.na(data10T))/(ncol(data10T)*nrow(data10T))
NA10T

dataNAT <- Xgen[which(is.na(rawdata$Term)), ]
NANA <- sum(is.na(dataNAT))/(ncol(dataNAT)*nrow(dataNAT))
NANA

NATerm <- c(NA1T, NA2T, NA3T, NA4T, NA5T, NA6T, NA7T, NA8T, NA9T, NA10T, NANA)

barplot(NATerm*100, main = 'Missings␣por␣trimestre', names.arg = c('02-1', '02-2',
                                                                   '02-3', '03-1',
                                                                   '03-2', '03-3',
                                                                   '04-1', '04-2',
                                                                   '04-3', '05-1',
                                                                   'NA'), las = 2,
        xlab = 'Trimestre', ylab = 'Porcentaje␣de␣missings', cex.main = 1,
        col = 'lightblue', cex.lab = 0.8, ylim = c(0, 80), cex.axis = 0.8)
abline(h = mean(NATerm*100), col = 'red', lty = 2)
legend(legend = 'Media', lty = 2, col = 'red', 'topleft', cex = 0.8)

gendata2 <- rbind(data1T, data2T)
gendata2 <- rbind(gendata2, data3T)
gendata2 <- rbind(gendata2, data4T)
gendata2 <- rbind(gendata2, data5T)
gendata2 <- rbind(gendata2, data6T)

rawdata2 <- rawdata[which(rawdata$Term == '02-1'),]
rawdata2 <- rbind(rawdata2, rawdata[which(rawdata$Term == '02-2'), ])
rawdata2 <- rbind(rawdata2, rawdata[which(rawdata$Term == '02-3'), ])
rawdata2 <- rbind(rawdata2, rawdata[which(rawdata$Term == '03-1'), ])
rawdata2 <- rbind(rawdata2, rawdata[which(rawdata$Term == '03-2'), ])
rawdata2 <- rbind(rawdata2, rawdata[which(rawdata$Term == '03-3'), ])

#write.csv(rawdata, file = "/media/blanca/7111-CB28/4 ESTADISTICA/TFG/data2.csv")

## PORCENTAJE DE DATOS FALTANTES POR VARIABLE
NAvar <- round(apply(rawdata2, 2, searchNA),2)
#NAvar[1:10]

hist(NAvar, main="Histograma␣del␣porcentaje␣de␣missings␣por␣variable",
     xlab='Porcentaje␣de␣missings', ylab="Numero␣de␣varibles", xlim = c(0, 100),
     ylim = c(0, 250), cex.main = 1, col = 'lightblue', las = 1, cex.lab = 0.8,
     cex.axis = 0.8,  cex.main = 1)

## PORCENTAJE DE DATOS FALTANTES POR INDIVIDUO
NAind <- round(apply(rawdata2, 1, searchNA),2)
#NAind[1:10]

hist(NAind, main='Histograma␣del␣porcentaje␣de␣missings␣por␣individuo',
     xlab='Porcentaje␣de␣missings', ylab="Numero␣de␣individuos", xlim = c(0, 100),
     ylim = c(0, 400), breaks = 10, cex.main = 1, col = 'lightblue', las = 1,
     cex.lab = 0.8, cex.axis = 0.8)

data2 <- rawdata2[which(NAind < 50), which(NAvar < 50)]

#write.csv(data, file = "/media/blanca/7111-CB28/4 ESTADISTICA/TFG/Adj/data.csv")
```

```
ndatos <- nrow(data2)*ncol(data2)
NAdata2 <- round(sum(is.na(data2))/ndatos*100, 2)

gendata2 <- data2[, 1:194]
ndatos <- nrow(gendata2)*ncol(gendata2)
NAgendata2 <- round(sum(is.na(gendata2))/ndatos*100, 2)

d1 <- dim(rawdata)-dim(data2)

d <- dim(data2)

## VI. ESTADISTICA DESCRIPTIVA
attach(data2)

### 6.1 VARIABLES DEMOGRAFICAS
#### GENERO
round(prop.table(table(Gender)), 2)

pie(prop.table(table(Gender)), labels = c('Femenino', 'Masculino'), cex = 0.8,
    col = c('lightblue2', 'mistyrose2'), main = 'Genero')
text(-0.1,0.4,'59%', cex = 0.8)
text(0.1,-0.35,'41%', cex = 0.8)

#### EDAD
par(cex.lab=0.8)
par(cex.axis=0.8)
#boxplot(data2$Age, main = 'Edad')
boxplot(Age ~ Gender, levels = c('Femenino', 'Masculino'), names =
          c('Femenino', 'Masculino'), cex.names = 0.8, las = 1, col = 'lightsteelblue1',
        main = 'Edad␣por␣genero', xlab = '', ylab = '')

#### RAZA
par(cex.lab=0.8)
par(cex.axis=0.8)
plot(droplevels(Race), ylim = c(0, 600), las = 2, names  = c('Africana', 'Asiatica',
                                                             'Caucasica', 'Hispanica',
                                                             'Otros'),
     col = 'lightblue', main = 'Raza')
text(0.75, 50, '4%', cex = 0.8)
text(2, 80, '9%', cex = 0.8)
text(3.15, 527, '78%', cex = 0.8)
text(4.3, 50, '4%', cex = 0.8)
text(5.55, 50, '4%', cex = 0.8)

table(Race)
round(prop.table(table(Race)), 2)

#### MANO DOMINANTE
pie(prop.table(table(Dom.Arm)), labels = c('Zurdos', 'Diestros'), cex = 0.8,
    col = c('lightblue2', 'mistyrose2'), main = 'Mano␣dominante')
text(0.5,0.1,'8%', cex = 0.8)
text(-0.4,0,'92%', cex = 0.8)
#round(prop.table(table(Dom.Arm)), 2)


### 6.2 VARIABLES DE RENDIMIENTO MUSCULAR
#### SECCION TRANSVERSAL BICEPS
bicD <- vector(length = nrow(data2))
bicND <- vector(length = nrow(data2))
```

```
for(i in 1:nrow(data2)){
  if(is.na(data2[, 'Dom.Arm'][i])){
    bicD[i] <- NA
    bicND[i] <- NA
  } else{
    if(data2[, 'Dom.Arm'][i] == 'L'){
      bicND[i] <- data2$Post_RBi_Avg[i] - data2$Pre_RBi_Avg[i]
      bicD[i] <- data2$Post_LBi_Avg[i] - data2$Pre_LBi_Avg[i]
    } else{
      bicND[i] <- data2$Post_LBi_Avg[i] - data2$Pre_LBi_Avg[i]
      bicD[i] <- data2$Post_RBi_Avg[i] - data2$Pre_RBi_Avg[i]
    }
  }
}

par(cex.lab=0.8)
par(cex.axis=0.8)
boxplot(bicND, bicD, names = c('Brazo␣no␣dominante', 'Brazo␣dominante'),
main = 'Diferencia␣despues␣y␣antes␣del␣entrenamiento
␣␣␣␣␣␣␣␣de␣la␣seccion␣transversal␣del␣biceps', cex.main = 1, las = 1,
    col = 'lightsteelblue1', cex.axis = 0.8)

var.test(bicND, bicD)

t.test(bicND, bicD, var.equal = FALSE, paired = TRUE)


#### SECCION TRANSVERSAL TRICEPS
triD<- vector(length = nrow(data2))
triND<- vector(length = nrow(data2))

for(i in 1:nrow(data2)){
  if(is.na(data2[, 'Dom.Arm'][i])){
    triD[i] <- NA
    triND[i] <- NA
  }else{
    if(data2[, 'Dom.Arm'][i] == 'L'){
      triND[i] <- data2$Post_RTri_Avg[i] - data2$Pre_RTri_Avg[i]
      triD[i] <- data2$Post_LTri_Avg[i] - data2$Pre_LTri_Avg[i]
    }else{
      triND[i] <- data2$Post_LTri_Avg[i] - data2$Pre_LTri_Avg[i]
      triD[i] <- data2$Post_RTri_Avg[i] - data2$Pre_RTri_Avg[i]
    }
  }
}

par(cex.lab=0.8)
par(cex.axis=0.8)
boxplot(triND, triD, names = c('Brazo␣no␣dominante', 'Brazo␣dominante'),
    main = 'Diferencia␣despues␣y␣antes␣del␣entrenamiento␣de␣la␣seccion␣transversal
␣␣␣␣del␣triceps',
    cex = 0.9, las = 1, cex.main = 1, col = 'lightsteelblue1', cex.axis = 0.8)

t.test(triND, triD, var.equal = TRUE, paired = TRUE)


#### TEST DE FUERZA ISOMETRICA
par(cex.lab=0.8)
```

```r
par(cex.axis=0.8)
boxplot(ND23_DIFF, D23_DIFF, names = c('Brazo no dominante', 'Brazo dominante'),
    main = 'Diferencia despues y antes del entrenamiento del test de fuerza',
    cex.main = 1, las = 1, col = 'lightsteelblue1', cex.axis = 0.8)

var.test(ND23_DIFF, D23_DIFF)

t.test(ND23_DIFF, D23_DIFF, var.equal = FALSE, paired = TRUE)


#### TEST DE REPETICION MAXIMA
par(cex.lab=0.8)
par(cex.axis=0.8)
boxplot(NDRM_DIFF, DRM_DIFF, names = c('Brazo no dominante', 'Brazo dominante'),
    main = 'Diferencia despues y antes del entrenamiento test de repeticion maxima',
    cex.main = 1, las = 1, col = 'lightsteelblue1', cex.axis = 0.8)

t.test(NDRM_DIFF, DRM_DIFF, var.equal = FALSE, paired = TRUE)


#### COMPARACION TEST DE GANANCIA DE FUERZA ISOMETRICA VS REPETICION MAXIMA
lattice::parallelplot(~ data2[, c('ND23_DIFF', 'NDRM_DIFF')] |Race, data = data2)


### 6.3 VARIABLES GENETICAS
#gendata <- data[, 1:196]
gendata <- gendata2

#### RECATEGORIZACION
recod <- function(v, separ){
  t <- table(unlist(strsplit(v, separ)))
  menor <- names(which.min(t))
  major <- names(which.max(t))
  homomenor <- paste(menor, menor, sep = separ)
  homomajor <- paste(major, major, sep = separ)
  heter1 <- paste(menor, major, sep = separ)
  heter2 <- paste(major, menor, sep = separ)
  v2 <- rep(NA, length(v))
  v2[homomenor == v] <- 2
  v2[homomajor == v] <- 0
  v2[heter1 == v] <- 1
  v2[heter2 == v] <- 1

  return(v2)
}

for (i in c(1:48, 50:142, 144:ncol(gendata))){
  gendata[, i] <- recod(as.character(gendata[, i]), '')
}

#gendata[, 'b2b'] <- recod(as.character(gendata[, 'b2b']), separ = ', ')
gendata[, 'b2b'][which(gendata[, 'b2b']=='xx')] <- NA
gendata[, 'b2b'][which(gendata[, 'b2b']=='XX')] <- NA
gendata[, 'b2b'] <- recod(as.character(gendata[, 'b2b']), separ = ', ')

v <- rep(NA, length(gendata[, 'pai1_4g5g']))
v[which(gendata[, 'pai1_4g5g'] == '4G4G')] <- 0
v[which(gendata[, 'pai1_4g5g'] == '4G5G')] <- 1
v[which(gendata[, 'pai1_4g5g'] == '5G5G')] <- 2
```

```
gendata[, 'pai1_4g5g'] <- v
#write.csv(gendata, file = "/home/blanca/Documents/Cole/4 ESTADISTICA/TFG/gendata.csv")

#### DEPURACION DE MISSINGS
gendataNA <- gendata
set.seed(2024)

for(i in 1:ncol(gendata)){
  n2 <- sum(gendata[, i] == 2, na.rm = TRUE)
  n1 <- sum(gendata[, i] == 1, na.rm = TRUE)
  n0 <- sum(gendata[, i] == 0, na.rm = TRUE)
  ntot <- n2 + n1 + n0


  p2 <- n2 / ntot
  p1 <- n1 / ntot
  p0 <- n0 / ntot

  nmis <- sum(is.na(gendata[, i]))

  pseudo <-sample(c(2, 1, 0), nmis, replace = TRUE, prob = c(p2, p1, p0))

  gendata[, i] [is.na(gendata[, i])] <- pseudo
}

v <- c()

for (i in 1:ncol(gendata)){
  n2 <- sum(gendata[, i] == 2, na.rm = TRUE)
  n1 <- sum(gendata[, i] == 1, na.rm = TRUE)
  n0 <- sum(gendata[, i] == 0, na.rm = TRUE)

  if(n2 == sum(!is.na(gendata[ ,i])) || n1 == sum(!is.na(gendata[ ,i])) ||
     n0 == sum(!is.na(gendata[ ,i]))){
    v <- c(v, i)
  }
}
gendata <- gendata[, -v]
gendataNA <- gendataNA[, -v]


#### FRECUENCIA DEL ALELO MENOS COMUN
gencau <- gendata[which(data2[, 'Race']=='Caucasian'),]
genaf <- gendata[which(data2[, 'Race']=='African␣Am'), ]
genas <- gendata[which(data2[, 'Race']=='Asian'), ]
genhisp <- gendata[which(data2[, 'Race']=='Hispanic'), ]
genoth <- gendata[which(data2[, 'Race']=='Other'), ]
gentot <- gendata

par(mfrow = c(3, 2))
menor <- vector()
for (i in 1:ncol(gencau)){
  n2 <- sum(gencau[, i] == 2)
  n1 <- sum(gencau[, i] == 1)
  n0 <- sum(gencau[, i] == 0)
  ntot <- n2 + n1 + n0

  pA <- ((2 * n2) + n1) / (2 * ntot)
  pB <- 1 - pA
```

```
  menor <- c(menor, min(pA, pB))
}

plot(density(menor), xlim = c(0, 0.5), main = 'MAF caucasicos', xlab = '',
    ylab = 'Densidad', las = 1)

#which.max(menor)


menor <- vector()
for (i in 1:ncol(genaf)){
  n2 <- sum(genaf[, i] == 2)
  n1 <- sum(genaf[, i] == 1)
  n0 <- sum(genaf[, i] == 0)
  ntot <- n2 + n1 + n0

  pA <- ((2 * n2) + n1) / (2 * ntot)
  pB <- 1 - pA

  menor <- c(menor, min(pA, pB))
}

plot(density(menor), xlim = c(0, 0.5), main = 'MAF afroamericanos', xlab = '',
    ylab = 'Densidad', las = 1)

#which.max(menor)



menor <- vector()
for (i in 1:ncol(genas)){
  n2 <- sum(genas[, i] == 2)
  n1 <- sum(genas[, i] == 1)
  n0 <- sum(genas[, i] == 0)
  ntot <- n2 + n1 + n0

  pA <- ((2 * n2) + n1) / (2 * ntot)
  pB <- 1 - pA

  menor <- c(menor, min(pA, pB))
}

plot(density(menor), xlim = c(0, 0.5),  main = 'MAF asiaticos', xlab = 'Frecuencia',
    ylab = 'Densidad', las = 1)

#which.max(menor)


menor <- vector()
for (i in 1:ncol(genhisp)){
  n2 <- sum(genhisp[, i] == 2)
  n1 <- sum(genhisp[, i] == 1)
  n0 <- sum(genhisp[, i] == 0)
  ntot <- n2 + n1 + n0

  pA <- ((2 * n2) + n1) / (2 * ntot)
  pB <- 1 - pA
```

```
  menor <- c(menor, min(pA, pB))
}

plot(density(menor), xlim = c(0, 0.5), main = 'MAF␣hispanicos', xlab = 'Frecuencia',
    ylab = 'Densidad', las = 1)

#which.max(menor)


menor <- vector()
for (i in 1:ncol(genoth)){
  n2 <- sum(genoth[, i] == 2)
  n1 <- sum(genoth[, i] == 1)
  n0 <- sum(genoth[, i] == 0)
  ntot <- n2 + n1 + n0

  pA <- ((2 * n2) + n1) / (2 * ntot)
  pB <- 1 - pA

  menor <- c(menor, min(pA, pB))
}

plot(density(menor), xlim = c(0, 0.5), main = 'MAF␣otros', xlab = 'Frecuencia',
    ylab = 'Densidad', las = 1)

#which.max(menor)

menor <- vector()
for (i in 1:ncol(gentot)){
  n2 <- sum(gentot[, i] == 2)
  n1 <- sum(gentot[, i] == 1)
  n0 <- sum(gentot[, i] == 0)
  ntot <- n2 + n1 + n0

  pA <- ((2 * n2) + n1) / (2 * ntot)
  pB <- 1 - pA

  menor <- c(menor, min(pA, pB))
}

plot(density(menor), xlim = c(0, 0.5), main = 'MAF␣total', xlab = 'Frecuencia',
    ylab = 'Densidad', las = 1, cex.axis = 0.8)

#which.max(menor)


#### LEY DE HARDY WEINBERG
library(HardyWeinberg)

#gencau <- gendata[which(data[, 201]=='Caucasian'), ]
#MakeCounts(gencau)[1:5,]

cauHW <- data.frame(gencau[, 1])
caupval <- vector()
for (i in 1:ncol(gencau)){
  hw <- HWExact((MakeCounts(gencau)[i,])[1:3], verbose = FALSE)
  caupval <- c(caupval, hw$pval)
  if(hw$pval > 0.05){
    cauHW <- data.frame(cauHW, gencau[, i])
```

```
  }
}
cauHW <- cauHW[, -1]

afHW <- data.frame(genaf[, 1])
afpval <- vector()
for (i in 1:ncol(genaf)){
  hw <- HWExact((MakeCounts(genaf)[i,])[1:3], verbose = FALSE)
  afpval <- c(afpval, hw$pval)
  if(hw$pval > 0.05){
    afHW <- data.frame(afHW, genaf[, i])
  }
}
afHW <- afHW[, -1]

asHW <- data.frame(genas[, 1])
aspval <- vector()
for (i in 1:ncol(genas)){
  hw <- HWExact((MakeCounts(genas)[i,])[1:3], verbose = FALSE)
  aspval <- c(aspval, hw$pval)
  if(hw$pval > 0.05){
    asHW <- data.frame(asHW, genas[, i])
  }
}
asHW <- asHW[, -1]

hispHW <- data.frame(genhisp[, 1])
hisppval <- vector()
for (i in 1:ncol(genhisp)){
  hw <- HWExact((MakeCounts(genhisp)[i,])[1:3], verbose = FALSE)
  hisppval <- c(hisppval, hw$pval)
  if(hw$pval > 0.05){
    hispHW <- data.frame(hispHW, genhisp[, i])
  }
}
hispHW <- hispHW[, -1]

othHW <- data.frame(genoth[, 1])
othpval <- vector()
for (i in 1:ncol(genoth)){
  hw <- HWExact((MakeCounts(genoth)[i,])[1:3], verbose = FALSE)
  othpval <- c(othpval, hw$pval)
  if(hw$pval > 0.05){
    othHW <- data.frame(othHW, genoth[, i])
  }
}
othHW <- othHW[, -1]

totHW <- data.frame(gentot[, 1])
totpval <- vector()
for (i in 1:ncol(gentot)){
  hw <- HWExact((MakeCounts(gentot)[i,])[1:3], verbose = FALSE)
  totpval <- c(totpval, hw$pval)
  if(hw$pval > 0.05){
    totHW <- data.frame(totHW, gentot[, i])
  }
}
totHW <- totHW[, -1]
```

```
#gendata2 <- sigHW

sigHWcau <- round((dim(gencau)[2]-dim(cauHW)[2])/dim(gencau)[2] * 100, 2)
sigHWaf <- round((dim(genaf)[2]-dim(afHW)[2])/dim(genaf)[2] * 100, 2)
sigHWas <- round((dim(genas)[2]-dim(asHW)[2])/dim(genas)[2] * 100, 2)
sigHWhisp <- round((dim(genhisp)[2]-dim(hispHW)[2])/dim(genhisp)[2] * 100, 2)
sigHWoth <- round((dim(genoth)[2]-dim(othHW)[2])/dim(genoth)[2] * 100, 2)
sigHWtot <- round((dim(gentot)[2]-dim(totHW)[2])/dim(gentot)[2] * 100, 2)

library(gap)

par(mfrow=c(3, 2))
qqunif(caupval, main = 'Dist. uniforme vs p-val caucasicos', col = 'blue4',
       las = 1, cex.axis = 0.8 )
qqunif(afpval, main = 'Dist. uniforme vs p-val afroamericanos', col = 'blue4',
       las = 1, cex.axis = 0.8)
qqunif(aspval, main = 'Dist. uniforme vs p-val asiaticos', col = 'blue4',
       las = 1, cex.axis = 0.8)
qqunif(hisppval, main = 'Dist. uniforme vs p-val hispanicos', col = 'blue4',
       las = 1, cex.axis = 0.8)
qqunif(othpval, main = 'Dist. uniforme vs p-val otros', col = 'blue4',
       las = 1, cex.axis = 0.8)
qqunif(totpval, main = 'Dist. uniforme vs p-val total', col = 'blue4',
       las = 1, cex.axis = 0.8)

par(mfrow=c(3, 2))

plot(-log10(caupval), ylim = c(0, max(-log10(caupval))), main = 'Distribucion
     p-valor caucasicos', xlab = 'SNPs', pch = 20, las = 1, ylab = '-log10(p-valor)',
     cex.axis = 0.8)
abline(h = -log10(0.05), col = 'dodgerblue3', lty = 2)
abline(h = -log10(0.05/193), col='red', lty = 2)
#abline(h = -log10(1-(1-0.05)^(1/193)), col='darkgreen', lty = 2)
legend('topleft', legend=c('v. critico 0.05', 'v. critico C.Bonferroni'),
       lty = 2, xpd = FALSE, col = c('dodgerblue3', 'red'), cex = 0.9)

plot(-log10(afpval), ylim = c(0, max(-log10(afpval))), main = 'Distribucion p-valor
     afroamericanos', xlab = 'SNPs', pch = 20, col = 'black', las = 1,
     ylab = '-log10(p-valor)', cex.axis = 0.8)
abline(h = -log10(0.05), col = 'dodgerblue3', lty = 2)
#abline(h = -log10(1-(1-0.05)^(1/193)), col='darkgreen', lty = 2)
abline(h = -log10(0.05/193), col='red', lty = 2)

plot(-log10(aspval), ylim = c(0, max(-log10(aspval))), main = 'Distribucion p-valor
     asiaticos',  xlab = 'SNPs', pch = 20, las = 1, ylab = '-log10(p-valor)',
    cex.axis = 0.8)
abline(h = -log10(0.05), col = 'dodgerblue3', lty = 2)
abline(h = -log10(1-(1-0.05)^(1/193)), col='darkgreen', lty = 2)
abline(h = -log10(0.05/193), col='red', lty = 2)

plot(-log10(hisppval), ylim = c(0, max(-log10(hisppval))), main = 'Distribucion
     p-valor hispanicos', xlab = 'SNPs', pch = 20, las = 1, ylab = '-log10(p-valor)',
     cex.axis = 0.8)
abline(h = -log10(0.05), col = 'dodgerblue3', lty = 2)
abline(h = -log10(1-(1-0.05)^(1/193)), col='darkgreen', lty = 2)
abline(h = -log10(0.05/193), col='red', lty = 2)


plot(-log10(othpval), ylim = c(0, max(-log10(othpval))), main = 'Distribucion p-valor
```

```
␣␣␣␣␣otros', xlab = 'SNPs', pch = 20, las = 1, ylab = '-log10(p-valor)', cex.axis = 0.8)
abline(h = -log10(0.05), col = 'dodgerblue3', lty = 2)
abline(h = -log10(1-(1-0.05)^(1/193)), col='darkgreen', lty = 2)
abline(h = -log10(0.05/193), col='red', lty = 2)

plot(-log10(totpval), ylim = c(0, max(-log10(totpval))), main = 'Distribucion␣p-valor
␣␣␣␣␣total', xlab = 'SNPs', pch = 20, las = 1, ylab = '-log10(p-valor)', cex.axis = 0.8)
abline(h = -log10(0.05), col = 'dodgerblue3', lty = 2)
#abline(h = -log10(1-(1-0.05)^(1/193)), col='darkgreen', lty = 2)
abline(h = -log10(0.05/193), col='red', lty = 2)

library(astsa)
FDR(caupval, 0.05)
FDR(afpval, 0.05)
FDR(aspval, 0.05)
FDR(hisppval, 0.05)
FDR(othpval, 0.05)
FDR(totpval, 0.05)

sum(caupval<=caupval[FDR(caupval, 0.05)])
sum(othpval<=othpval[FDR(othpval, 0.05)])
sum(totpval<=totpval[FDR(totpval, 0.05)])

gendata <- gendata[ , -which.min(caupval)]

### 6.4 OTRAS
### CENTRO
par(cex.lab=0.8)
par(cex.axis=0.8)
plot(droplevels(Center), ylim = c(0, 120), las = 1, cex.main = 0.9, col = 'lightblue',
     cex.axis = 0.8, xlab = 'Centro', ylab = 'Numero␣de␣individuos',
     main = 'Numero␣de␣individuos␣por␣centro')
text(0.75, 69, '10%', cex = 0.8)
text(1.95, 84, '12%', cex = 0.8)
text(3.15, 68.5, '10%', cex = 0.8)
text(4.3, 88, '13%', cex = 0.8)
text(5.55, 95, '14%', cex = 0.8)
text(6.75, 111, '16%', cex = 0.8)
text(7.95, 98.65, '14%', cex = 0.8)
text(9.15, 79, '11%', cex = 0.8)

round(prop.table(table(Center)), 2)

#### TRIMESTRE
par(cex.lab=0.8)
par(cex.axis=0.8)
plot(droplevels(Term), ylim = c(0, 200), cex.main = 0.8, las = 1, col = 'lightblue',
     xlab = 'Timestre', ylab = 'Numero␣de␣individuos', cex.axis = 0.8,
     main = 'Numero␣de␣individuos␣por␣trimestre')
text(0.75, 153, '22%', cex = 0.8)
text(1.95, 79, '11%', cex = 0.8)
text(3.15, 140, '20%', cex = 0.8)
text(4.3, 170, '25%', cex = 0.8)
text(5.55, 49.5, '6%', cex = 0.8)
text(6.75, 111, '16%', cex = 0.8)

round(prop.table(table(droplevels(Term))), 2)
```

```
### VII. EXPLORACION DATOS GENETICOS
NAvargen <- round(apply(gendataNA, 2, searchNA),2)
NAindgen <- round(apply(gendataNA, 1, searchNA),2)

mean(NAvargen)
mean(NAindgen)

plot(1:193, NAvargen, main = 'Grafico␣SNP␣vs␣porcentaje␣de␣missings', xlab = 'SNPs',
     ylim = c(0, 50), ylab = 'Porcentaje␣de␣missings', cex.main = 1, las = 1, pch = 20,
     col = 'black', cex.lab = 0.8, cex.axis = 0.8)
abline(h = mean(NAvargen), col = 'red', lty = 2)
legend(legend='Media', 'topleft', lty = 2, col = 'red', cex = 0.8)

# col  <- c(rep('black',143), rep('red', 70), rep('blue', 130), rep('green', 160),
#           rep('pink', 40), rep('yellow', 102))
# plot(1:645, NAindgen, main = 'Grafico individuos vs porcentaje de missings',
#      xlab = 'Individuos', ylab = 'porcentaje de missings', pch = 19, col = col)
# abline(h = mean(NAindgen), col = 'red', lty = 2)

#for (i in 1:ncol(gendata)){
#  gendata[, i][which(is.na(gendata[, i]))] <- colMeans(gendata, na.rm = TRUE)[i]
#}


### 7.4 ANALISIS DE COMPONENTES PRINCIPALES
library(stats)
gendata <- as.matrix(gendata)
genPCA <- princomp(gendata, cor=FALSE)
#summary(genPCA)
#biplot(genPCA)

prop_varianza <- genPCA$sdev^2 / sum(genPCA$sdev^2)
barplot(prop_varianza[1:10]*100, las = 2, ylim = c(0, 5), main = 'Varianza␣explicada
␣␣␣␣␣␣␣␣por␣componente', ylab = 'Porcentaje␣de␣varianza␣explicada', names.arg =
          c('Comp.␣1', 'Comp.␣2','Comp.␣3','Comp.␣4','Comp.␣5','Comp.␣6','Comp.␣7',
            'Comp.␣8','Comp.␣9','Comp.␣10'), col = 'lightblue', cex.lab = 0.8,
        cex.axis = 0.8)

par(pty="s", xpd = TRUE, mfrow = c(1,1))
#opar <- par(pty = 's')
#par(opar)
Fp <- genPCA$scores
#genPCA$sdev
#genPCA$rotation
#genPCA$center
#genPCA$x
col <- rep('gold', 645)
col[which(data2[, 'Race']=='Caucasian')] <- 'brown2'
col[which(data2[, 'Race']=='African␣Am')] <- 'royalblue3'
col[which(data2[, 'Race']=='Hispanic')] <- 'limegreen'
col[which(data2[, 'Race']=='Other')] <- 'black'
plot(Fp[,1], Fp[,2], col=col, asp = 1, pch = 20, cex = 0.8, main = 'Primera␣y␣segunda
␣␣␣␣␣componente', xlab = 'Primera␣componente', ylab ='Segunda␣componente', las = 1,
     cex.lab = 0.7, cex.main = 0.9, cex.axis = 0.8)
legend(legend = c('Asiaticos', 'Caucasicos', 'Afroamericanos',  'Hispanicos', 'Otros'),
       col = c('gold', 'brown2', 'royalblue3', 'limegreen', 'black'), 'topright',
       pch = 20, cex = 0.7, inset=c(-0.58,0.35))

par(pty="s", xpd = TRUE, mfrow = c(1,2))
```

```
plot(Fp[,1], Fp[,3], col=col, asp = 1, pch = 20, cex = 0.85, main = 'Primera y tercera
componente', xlab = 'Primera componente', ylab ='Tercera componente', las = 1,
     cex.lab = 0.8)
#legend(legend = c('Asiaticos', 'Caucasicos', 'Afroamericanos',  'Hispanicos', 'Otros'),
#       col = c('gold', 'brown2', 'royalblue3', 'limegreen', 'black'), 'topright',
#       pch = 20,cex = 0.8, inset=c(-0.55,0.35))
plot(Fp[,2], Fp[,3], col=col, asp = 1, pch = 20, cex = 0.85, main = 'Segunda y tercera
componente',  xlab = 'Segunda componente', ylab ='Tercera componente', las = 1,
     cex.lab = 0.8)
#legend(legend = c('Asiaticos', 'Caucasicos', 'Afroamericanos', 'Hispanicos', 'Otros'),
#       col = c('gold', 'brown2', 'royalblue3', 'limegreen', 'black'), 'topright',
#       pch = 20, cex = 0.8, inset=c(-0.55,0.35))

## VIII. MODELOS ESTADISTICOS DE RENDIMIENTO MUSCULAR
set.seed(2024)
par(mfrow=c(2,2))
ad <- c(rnorm(100, 0, 2), rnorm(100, 5, 2), rnorm(100, 10, 2))
ad2 <- as.factor(c(rep('AA', 100), rep('AB', 100), rep('BB', 100)))
boxplot(ad ~ ad2, main = 'Esquema aditivo', las = 1, xlab = 'Niveles SNP',
        ylab = 'ND23_DIFF', cex.main = 1,
        cex.lab = 0.9, cex.axis = 0.9, col = 'lightsteelblue1')
points(x = 1:3,  y = (tapply(ad, ad2, median, na.rm = TRUE)),
       pch = 19, col = 'red', type = 'b', lty = 2)

rec <- c(rnorm(100, 0, 2), rnorm(100, 0, 2), rnorm(100, 10, 2))
rec2 <- as.factor(c(rep('AA', 100), rep('AB', 100), rep('BB', 100)))
boxplot(rec ~ rec2, main = 'Esquema recesivo', las = 1, xlab = 'Niveles SNP',
        ylab = 'ND23_DIFF', cex.main = 1,
        cex.lab = 0.9, cex.axis = 0.9, col = 'lightsteelblue1')
points(x = 1:3,  y = (tapply(rec, rec2, median, na.rm = TRUE)),
       pch = 19, col = 'red', type = 'b', lty = 2)

dom <- c(rnorm(100, 0, 2), rnorm(100, 10, 2), rnorm(100, 10, 2))
dom2 <- as.factor(c(rep('AA', 100), rep('AB', 100), rep('BB', 100)))
boxplot(dom ~ dom2, main = 'Esquema dominante', las = 1, xlab = 'Niveles SNP',
        ylab = 'ND23_DIFF', cex.main = 1,
        cex.lab = 0.9, cex.axis = 0.9, col = 'lightsteelblue1')
points(x = 1:3,  y = (tapply(dom, dom2, median, na.rm = TRUE)),
       pch = 19, col = 'red', type = 'b', lty = 2)

codom <- c(rnorm(100, 0, 2), rnorm(100, 5, 2), rnorm(100, 0, 2))
codom2 <- as.factor(c(rep('AA', 100), rep('AB', 100), rep('BB', 100)))
boxplot(codom ~ codom2, main = 'Esquema codominante', las = 1, xlab = 'Niveles SNP',
        ylab = 'ND23_DIFF', cex.main = 1,
        cex.lab = 0.9, cex.axis = 0.9, col = 'lightsteelblue1', ylim = c(-4, 12))
points(x = 1:3,  y = (tapply(codom, codom2, median, na.rm = TRUE)),
       pch = 19, col = 'red', type = 'b', lty = 2)
codom_2 <- c(rnorm(100, 5, 2), rnorm(100, 0, 2), rnorm(100, 5, 2))
codom2_2 <- as.factor(c(rep('AA', 100), rep('AB', 100), rep('BB', 100)))
boxplot(codom_2 ~ codom2_2, main = 'Esquema codominante', las = 1, xlab = 'Niveles SNP',
        ylab = 'ND23_DIFF', cex.main = 1,
        cex.lab = 0.9, cex.axis = 0.9,  add = TRUE, lty = 3)
points(x = 1:3,  y = (tapply(codom_2, codom2_2, median, na.rm = TRUE)),
       type = 'b', lty = 3, pch = 20, cex = 0.9)


### 8.1 GANANCIA DE FUERZA ISOMETRICA
plot(ND23_DIFF)
```

```
#### SELECCION DEL MODELO
library(MASS)

#resp <- data2$ND23_DIFF

#lm1 <- lm(resp ~ Center * Term * Gender * Age * Race * Pre.weight * Pre.height *
#pre.BMI * SBP * DBP * FLGU * TG * CHOL * HDL_C * CHOL_HDL_C * VLDL_TG * LDL_C *
#FINS * CRP * HOMA * Met_syn, data2)

#lm2<-lm(resp ~  Center * Term * Gender * Age * Race * Pre.weight * Pre.height *
#pre.BMI * SBP * DBP * FLGU * TG * CHOL * HDL_C * CHOL_HDL_C * VLDL_TG * LDL_C *
#FINS * CRP *  HOMA * Met_syn, data=lm1$model)

#stepAIC(lm2, direction = 'both')

ND23_DIFF <- data2$ND23_DIFF

lm1 <- lm(ND23_DIFF ~  Center + Term + Gender + Age + Race + Pre.weight +
            Pre.height + pre.BMI + SBP + DBP + FLGU + TG + CHOL + HDL_C +
            CHOL_HDL_C + VLDL_TG + LDL_C + FINS + CRP + HOMA + Met_syn,
          data2)

lm2<-lm(ND23_DIFF ~  Center + Term + Gender + Age + Race + Pre.weight +
          Pre.height + pre.BMI + SBP + DBP + FLGU + TG + CHOL + HDL_C +
          CHOL_HDL_C + VLDL_TG + LDL_C + FINS + CRP +
          HOMA + Met_syn, data=lm1$model)

#stepAIC(lm2, direction = 'both')

lmND23_DIFF <- lm(ND23_DIFF ~ Center + Term + Gender + DBP + VLDL_TG + Race, data2)
summary(lmND23_DIFF)

# lmND23_DIFF_int <- lm(resp ~ Center + Term + Gender + DBP + VLDL_TG + Race +
#Center:Term + Center:Gender + Center:DBP + Center:VLDL_TG + Term:Gender + Term:DBP +
#Term:VLDL_TG  + Gender:DBP + Gender:VLDL_TG  + DBP:VLDL_TG + Race:Center + Race:Term +
#Race:DBP, data2)

# summary(lmND23_DIFF_int)

#stepAIC(lmND23_DIFF_int, direction = 'both')

lm0 <- lm(ND23_DIFF ~ Center + Term + Gender + DBP + VLDL_TG + Race, data2)
lm1 <- lm(ND23_DIFF ~ Term + Gender + DBP + VLDL_TG + Race, data2)
anova(lm0, lm1)

lm0 <- lm(ND23_DIFF ~ Center + Term + Gender + DBP + VLDL_TG + Race, data2)
lm1 <- lm(ND23_DIFF ~ Center + Gender + DBP + VLDL_TG + Race, data2)
anova(lm0, lm1)




##### MODELO ADITIVO
valorp <- numeric()
w <- c()

for (i in 1:ncol(gendata)){
  w <- gendata[, i]
  p <- summary(lm(ND23_DIFF ~ as.numeric(w) + Center + Term + Gender + DBP + VLDL_TG +
```

```
                    Race, data = data2))
  valorp[i] <- p$coefficients[2,4]
}

qqunif(valorp, main = 'Q-Q Plot p-valores vs. distribucion uniforme', col = 'blue4',
       cex.lab = 0.8,
       las = 1, cex.main = 1, cex.axis = 0.8)

par(mfrow = c(1,2))
plot(-log10(valorp), ylim=c(0, 5), main = '-log10(p-val)', pch = 20, col = 'black',
     cex.lab = 0.8, xlab = 'SNPs', ylab = '-log10(p-valor)', las = 1, cex.main = 1,
     cex.axis = 0.8)
abline(h = -log10(0.05), col = 'dodgerblue3', lty = 2)
abline(h = -log10(0.05/193), col='red', lty = 2)
legend(legend = c('v. critico 0.05', 'v. critico C.Bonferroni'), col =
         c('dodgerblue3', 'red'), lty = 2, 'topright', cex = 0.9)
fdr <- p.adjust(valorp, method = 'fdr')
plot(-log10(fdr),  ylim=c(0, 5), main = '-log10(p-valor) ajustados por el metodo FDR',
     pch = 20, col = 'black', cex.lab = 0.8, xlab = 'SNPs', ylab =
       '-log10(p-valor) ajustados con FDR', las = 1, cex.main = 1, cex.axis = 0.8 )
abline(h = -log10(0.05), col = 'dodgerblue3', lty = 2)
legend(legend = 'v. critico 0.05', col = 'dodgerblue3', lty = 2,
       'topright', cex = 0.85)


FDR(valorp, 0.05)

p <- colnames(gendata[,order(-log10(valorp), decreasing = TRUE)])[1:20]
p_2 <- round(sort(valorp, decreasing = FALSE), 3)[1:20]
p_p <- c("adrb2 1042713", "resistin g540a", "kchj11 rs5219", "vdr bsm1",  "rs11630261",
         "gs s287nga","akt1 c832g c3359g", "resistin c980g", "akt1 g4362c", "igf1 t1245c",
         "carp c105t", "vim 6602186", "carp a8470g", "vdr taq1", "igfbp3 2132570",
         "myod1 rs2249104", "cast rs754615", "resistin c398t", "ankrd6 p636l",
         "akt1 g1780a g363a")

par(mfrow = c(1,2))

col <- rep('dimgrey', length(ND23_DIFF))
col[gendata[, p[1]] == 1] <- 'skyblue3'
col[gendata[, p[1]] == 2] <- 'darksalmon'
plot(ND23_DIFF, col = col, pch = 20, xlab = 'Individuos', main = 'ND23_DIFF segun
     "adrb2_1042713"', cex.main = 1, las = 1, cex.lab = 0.8, cex.axis = 0.8)
abline(h = mean(ND23_DIFF, na.rm = TRUE), lty = 2, col = 'red', lwd = 2)
legend(legend = c('GG', 'GA', 'AA'), col = c('dimgrey', 'skyblue3', 'darksalmon'),
       pch = 20, 'topright', cex = 0.8)
legend(legend = 'Media', lty = 2, col = 'red', cex = 0.8, 'topleft', lwd = 2)

adrb2_1042713 <- gendata[ , which(colnames(gendata) == p[1])]

boxplot(ND23_DIFF ~ gendata[, p[1]], main = c('adrb2_1042713 ~ ND23_DIFF'),
        xlab = p[1],  ylab = 'ND23_DIFF', axes = FALSE, ylim = c(-50, 100),
        cex.main = 1, cex.lab = 0.8, cex.axis = 0.8, col = 'lightsteelblue1')
axis(1, at = c(1,2,3), labels = c('GG', 'GA', 'AA'), cex.axis = 0.8)
axis(2, at = c(-50,0,50,100),las = 1, cex.axis = 0.8)
points(x = 1:3,  y = (tapply(ND23_DIFF, gendata[,p[1]], mean, na.rm = TRUE)),
       pch = 19, col = 'red', type = 'b', lty = 2)
legend(legend = 'Media', 'topright', pch = 19, col = 'red', cex = 0.8)

lm_ND23_DIFF_A <- lm(ND23_DIFF ~ as.numeric(adrb2_1042713) + Center + Term +
```

```
                        Gender + DBP + VLDL_TG + Race,  data2)
sA <- summary(lm_ND23_DIFF_A)
sA

par(mfrow =c (1, 2))
plot(lm_ND23_DIFF_A, which = 1:2, las = 1, cex.lab = 0.8, cex.axis = 0.8)

ND23_DIFF_prov <- ND23_DIFF
ND23_DIFF <- ND23_DIFF[-328]

adrb2_1042713_prov <- adrb2_1042713
adrb2_1042713 <- adrb2_1042713[-328]

lm_ND23_DIFF_A <- lm(ND23_DIFF ~ as.numeric(adrb2_1042713) + Center + Term +
                        Gender + DBP + VLDL_TG + Race,
                      data2[-328,])
sA <- summary(lm_ND23_DIFF_A)
sA

ND23_DIFF <- ND23_DIFF_prov
adrb2_1042713 <- adrb2_1042713_prov


##### MODELO RECESIVO
valorp <- numeric()
w <- c()

for (i in 1:ncol(gendata)){
  wb <- gendata[, i]
  wb[wb == 1] <- 0
  wb[wb == 2] <- 1

  p <- summary(lm(ND23_DIFF ~ as.numeric(wb) + Center + Term + Gender +
                    DBP + VLDL_TG + Race,
                  data = data2))
  valorp[i] <- p$coefficients[2,4]
}

qqunif(valorp, main = 'Q-Q␣Plot␣p-valores␣vs.␣distribucion␣uniforme',
       col = 'blue4', cex.lab = 0.8,
       las = 1, cex.main = 1, cex.axis = 0.8)

par(mfrow = c(1,2))
plot(-log10(valorp), ylim=c(0, 5), main = '-log10(p-val)', pch = 20,
     col = 'black', cex.lab = 0.8, xlab = 'SNPs', ylab = '-log10(p-valor)',
     las = 1, cex.main = 1, cex.axis = 0.8)
abline(h = -log10(0.05), col = 'dodgerblue3', lty = 2)
abline(h = -log10(0.05/193), col='red', lty = 2)
legend(legend = c('v.␣critico␣0.05', 'v.␣critico␣C.Bonferroni'),
       col = c('dodgerblue3', 'red'), lty = 2, 'topright', cex = 0.85)
fdr <- p.adjust(valorp, method = 'fdr')
plot(-log10(fdr),  ylim=c(0, 5), main = '-log10(p-valor)␣ajustados␣por␣el␣metodo␣FDR',
     pch = 20, col = 'black', cex.lab = 0.8, xlab = 'SNPs', ylab =
       '-log10(p-valor)␣ajustados␣con␣FDR', las = 1, cex.main = 1, cex.axis = 0.8 )
abline(h = -log10(0.05), col = 'dodgerblue3', lty = 2)
legend(legend = 'v.␣critico␣0.05', col = 'dodgerblue3', lty = 2,
       'topright', cex = 0.85)

FDR(valorp, 0.05)
```

```
p2 <- colnames(gendata[,order(-log10(valorp), decreasing = TRUE)])[1:15]
p2_2 <- round(sort(valorp, decreasing = FALSE), 3)[1:15]
p2_p <- c("vdr␣taq1", "ankrd6␣q122e", "pcr5␣snp1", "vdr␣bsm1", "adrb2␣1042713",
          "vdr␣rs731236", "nr3c1␣rs10482614", "cast␣rs7724759", "mgst3␣4147542",
          "fbox32␣rs3739287", "akt1␣c9756a␣c11898t", "akt1␣g4362c",
          "tcfl72␣rs12255372", "bcl6␣4686467", "tcfl72␣12255372")

par(mfrow = c(1,2))

col <- rep('dimgrey', length(gendata[,p2[1]]))
col[gendata[, p2[1]] == 1] <- 'skyblue3'
col[gendata[, p2[1]] == 2] <- 'darksalmon'
plot(ND23_DIFF, col = col, pch = 20, ylim = c(-50, 100), main = 'ND23_DIFF␣segun
␣␣␣␣␣"vdr_taq1"', ylab = 'ND23_DIFF', las = 1,
     cex.lab = 0.8, xlab = 'Individuos', cex.main = 1, cex.axis = 0.8)
abline(h = mean(ND23_DIFF, na.rm = TRUE), lty = 2, col = 'red', lwd = 2)
legend(legend = c('TT', 'TC', 'CC'), col = c('dimgrey', 'skyblue3', 'darksalmon'),
       pch = 20, 'topright', cex = 0.65)
legend(legend = 'Media', lty = 2, col = 'red', cex = 0.65, 'topleft', lwd = 2)

w <- gendata[ , p2[1]]
boxplot(ND23_DIFF ~ gendata[, p2[1]], main = c('vdr_taq1~␣ND23_DIFF'),
        xlab = p2[1], ylab = 'ND23_DIFF', axes = FALSE, ylim = c(-50, 100),
        cex.lab = 0.8, cex.main = 1, col = 'lightsteelblue1')
axis(1, at = c(1,2,3), labels = c('TT', 'TC', 'CC'), cex.axis = 0.8)
axis(2, at = c(-50,0,50,100), las = 1, cex.axis = 0.8)
points(x = 1:3,  y = (tapply(ND23_DIFF, gendata[,p2[1]], mean,
                             na.rm = TRUE)),  pch = 19, col = 'red', type = 'b',
       lty = 2)
legend(legend = 'Media', 'topright', pch = 19, col = 'red', cex = 0.8)

vdr_taq1b <- gendata[, 'vdr_taq1']
vdr_taq1b[gendata[, 'vdr_taq1'] == 1] <- 0
vdr_taq1b[gendata[, 'vdr_taq1'] == 2] <- 1

lm_ND23_DIFF_B <- lm(ND23_DIFF ~ as.numeric(vdr_taq1b) + Center + Term + Gender
                     + DBP + VLDL_TG + Race, data2)
sB <- summary(lm_ND23_DIFF_B)
sB

par(mfrow =c (1, 2))
plot(lm_ND23_DIFF_B, which = 1:2, las = 1, cex.lab = 0.8, cex.axis = 0.8)

ND23_DIFF_prov <- ND23_DIFF
ND23_DIFF <- ND23_DIFF[-328]

vdr_taq1_prov <- vdr_taq1b
vdr_taq1b <- vdr_taq1b[-328]

lm_ND23_DIFF_B <- lm(ND23_DIFF ~ as.numeric(vdr_taq1b) + Center + Term +
                       Gender + DBP + VLDL_TG + Race,  data2[-328,])
sB <- summary(lm_ND23_DIFF_B)
sB

ND23_DIFF <- ND23_DIFF_prov
vdr_taq1b <- vdr_taq1_prov
```

```
##### MODELO DOMINNTE
valorp <- numeric()
w <- c()

for (i in 1:ncol(gendata)){
  wc <- gendata[, i]
  wc[wc == 2] <- 1

  p <- summary(lm(ND23_DIFF ~ as.numeric(wc) + Center + Term + Gender +
                    DBP + VLDL_TG + Race, data = data2))
  valorp[i] <- p$coefficients[2,4]
}

qqunif(valorp, main = 'Q-Q Plot p-valores vs. distribucion uniforme',
       col = 'blue4', cex.lab = 0.8,las = 1, cex.main = 1, cex.axis = 0.8)

par(mfrow = c(1,2))
plot(-log10(valorp), ylim=c(0, 5), main = '-log10(p-val)', pch = 20,
     col = 'black', cex.lab = 0.8, xlab = 'SNPs', ylab = '-log10(p-valor)',
     las = 1, cex.main = 1, cex.axis = 0.8)
abline(h = -log10(0.05), col = 'dodgerblue3', lty = 2)
abline(h = -log10(0.05/193), col='red', lty = 2)
legend(legend = c('v. critico 0.05', 'v. critico C.Bonferroni')
       , col = c('dodgerblue3', 'red'), lty = 2, 'topright', cex = 0.9)
fdr <- p.adjust(valorp, method = 'fdr')
plot(-log10(fdr),  ylim=c(0, 5), main = '-log10(p-valor) ajustados por el metodo FDR',
     pch = 20, col = 'black', cex.lab = 0.8, xlab = 'SNPs', ylab =
       '-log10(p-valor) ajustados con FDR', las = 1, cex.main = 1, cex.axis = 0.8 )
abline(h = -log10(0.05), col = 'dodgerblue3', lty = 2)
legend(legend = 'v. critico 0.05', col = 'dodgerblue3', lty = 2,
       'topright', cex = 0.85)

FDR(valorp, 0.05)

p3 <- colnames(gendata[,order(-log10(valorp), decreasing = TRUE)])[1:10]
p3_2 <- round(sort(valorp, decreasing = FALSE), 3)[1:10]
p3_p <- c("resistin g540a", "igf1 t1245c", "kchj11 rs5219", "carp c105t",
          "akt1 c832g c3359g", "carp a8470g", "cast rs754615", "resistin c980g",
          "rs11630261", "gs s287nga")

par(mfrow = c(1,2))

col <- rep('dimgrey', length(ND23_DIFF))
col[gendata[, p3[1]] == 1] <- 'skyblue3'
col[gendata[, p3[1]] == 2] <- 'darksalmon'
plot(ND23_DIFF, col = col, pch = 20, lwd = 2, main = 'ND23_DIFF segun
     "resistin_g540a"', cex.main = 1, las = 1, cex.main = 1,
     cex.lab = 0.9, cex.axis = 0.9, ylim = c(-50, 150))
abline(h = mean(ND23_DIFF, na.rm = TRUE), lty = 2, col = 'red')
legend(legend = c('GG', 'GA', 'AA'), col = c('dimgrey', 'skyblue3',
                                             'darksalmon'), pch = 20,
       'topright', cex = 0.9)
legend(legend = 'Media', lty = 2, col = 'red', cex = 0.8, 'topleft',
       lwd = 2)

boxplot( ND23_DIFF ~ gendata[, p3[1]], main = c('"resistin_g540a" ~
          ND23_DIFF'), xlab = p3[1],
         ylab = 'ND23_DIFF', axes = FALSE, cex.main = 1, cex.lab = 0.8,
         cex.axis = 0.8, ylim = c(-50, 150), col = 'lightsteelblue1')
```

```
axis(1, at = c(1,2,3), labels = c('GG','GA','AA'), cex.axis = 0.8)
axis(2, las = 1, cex.axis = 0.8)
points(x = 1:3,  y = (tapply(ND23_DIFF, gendata[,p3[1]], mean, na.rm = TRUE)),
       pch = 19, col = 'red', type = 'b', lty = 2)
legend(legend = 'Media', 'topright', pch = 19, col = 'red', cex = 0.9)

resistin_g540ac <- gendata[, 'resistin_g540a']
resistin_g540ac[gendata[, 'resistin_g540a'] == 2] <- 1

lm_ND23_DIFF_C <- lm(ND23_DIFF ~ as.numeric(resistin_g540ac) + Center +
                       Term + Gender + DBP + VLDL_TG + Race,
                     data2)
sB <- summary(lm_ND23_DIFF_C)
sB

par(mfrow =c (1, 2))
plot(lm_ND23_DIFF_C, which = 1:2, las = 1, cex.lab = 0.8, cex.axis = 0.8)

ND23_DIFF <- ND23_DIFF[-328]

resistin_g540ac_prov <- resistin_g540ac
resistin_g540ac <- resistin_g540ac[-328]

lm_ND23_DIFF_C <- lm(ND23_DIFF ~ as.factor(resistin_g540ac) + Center +
                       Term + Gender + DBP + VLDL_TG + Race,  data2[-328, ])
sC <- summary(lm_ND23_DIFF_C)
sC

ND23_DIFF <- ND23_DIFF_prov
resistin_g540ac <- resistin_g540ac_prov


##### MODELO CODOMINANTE
valorp1 <- numeric()
w <- c()

for (i in 1:ncol(gendata)){
  w <- as.factor(gendata[, i])
  lm0 <- lm(ND23_DIFF ~ w + Center + Term + Gender + DBP + VLDL_TG + Race,
            data = data2)
  lm1 <- lm(ND23_DIFF ~ Center + Term + Gender + DBP + VLDL_TG + Race,
            data = data2)
  a <- anova(lm0,lm1)
  valorp1 <- c(valorp1, a$`Pr(>F)`[2])
}

qqunif(valorp1, main = 'Q-Q␣Plot␣p-valores␣vs.␣dist␣uniforme',
       col = 'blue4', las = 1,  cex.lab = 0.8, cex.main = 1, cex.axis = 0.8)

par(mfrow = c(1,2))
plot(-log10(valorp1), ylim=c(0, 5), main = '-log10(p-valor)', pch = 20,
     las = 1, xlab = 'SNPs', ylab = '-log10(-pvalor)', cex.lab = 0.8,
     cex.main = 1, cex.axis = 0.8)
abline(h = -log10(0.05), col = 'dodgerblue3', lty = 2)
abline(h = -log10(0.05/193), col='red', lty = 2)
legend(legend = c('v.␣critico␣0.05', 'v.␣critico␣C.Bonferroni'),
       col = c('dodgerblue3', 'red'), lty = 2, 'topright', cex = 0.85)
fdr <- p.adjust(valorp1, method = 'fdr')
plot(-log10(fdr),  ylim=c(0, 5), main = '-log10(p-valor)␣ajustados␣por␣el␣metodo␣FDR',
```

```
     pch = 20, col = 'black', cex.lab = 0.8, xlab = 'SNPs', ylab =
       '-log10(p-valor) ajustados con FDR', las = 1, cex.main = 1, cex.axis = 0.8 )
abline(h = -log10(0.05), col = 'dodgerblue3', lty = 2)
legend(legend = 'v. critico 0.05', col = 'dodgerblue3', lty = 2,
       'topright', cex = 0.85)

FDR(valorp, 0.05)

p4 <- colnames(gendata[,order(-log10(valorp1), decreasing = TRUE)])[1:20]
p4_2 <- round(sort(valorp1, decreasing = FALSE)[1:20], 3)
p4_p <- c("vdr taq1", "pcr5 snp1" , "ankrd6 q122e", "vdr bsm1", "adrb2 1042713",
          "vdr rs731236", "cast rs7724759", "mgst3 4147542", "fbox32 rs3739287",
          "resistin g540a", "nr3c1 rs10482614", "tcfl72 7903146", "igf1 t1245c",
          "kchj11 rs5219", "lepr 1137100", "fst 722910", "tcfl72 rs7903146",
          "gs s287nga", "carp c105t", "akt1 c832g c3359g")

par(mfrow = c(1,2))

col <- rep('dimgrey', length(gendata[,p4[1]]))
col[gendata[, p4[1]] == 1] <- 'skyblue3'
col[gendata[, p4[1]] == 2] <- 'darksalmon'
plot(ND23_DIFF, col = col, pch = 20, ylim = c(-50, 100), main = 'ND23_DIFF segun
     "vdr_taq1"', ylab = 'ND23_DIFF', las = 1, cex.lab = 0.8, xlab = 'Individuos',
     cex.main = 1, cex.axis = 0.8)
abline(h = mean(ND23_DIFF, na.rm = TRUE), lty = 2, col = 'red', lwd = 2)
legend(legend = c('TT', 'TC', 'CC'), col = c('dimgrey', 'skyblue3', 'darksalmon'),
       pch = 20, 'topright', cex = 0.65)
legend(legend = 'Media', lty = 2, col = 'red', cex = 0.65, 'topleft', lwd = 2)

w <- gendata[ , p4[1]]
boxplot(data2$ND23_DIFF ~ gendata[, p4[1]], main = c('vdr_taq1 ~ ND23_DIFF'),
        col = 'lightsteelblue1',
        xlab = p4[1], ylab = 'ND23_DIFF', axes = FALSE, ylim = c(-50, 100),
        cex.lab = 0.8, cex.main = 1)
axis(1, at = c(1,2,3), labels = c('TT', 'TC', 'CC'), cex.axis = 0.8)
axis(2, at = c(-50,0,50,100), las = 1, cex.axis = 0.8)
points(x = 1:3,  y = (tapply(ND23_DIFF, gendata[,p4[1]], mean,
                             na.rm = TRUE)),  pch = 19, col = 'red',
       type = 'b', lty = 2)
legend(legend = 'Media', 'topright', pch = 19, col = 'red', cex = 0.8)

lm_ND23_DIFF_D <- lm(ND23_DIFF ~ as.factor(vdr_taq1) + Center + Term + Gender +
                       DBP + VLDL_TG + Race, data2)
sD <- summary(lm_ND23_DIFF_D)
sD

par(mfrow =c (1, 2))
plot(lm_ND23_DIFF_D, which = 1:2, las = 1, cex.lab = 0.8, cex.axis = 0.8)

ND23_DIFF_prov <- ND23_DIFF
ND23_DIFF <- ND23_DIFF[-328]

vdr_taq1_prov <- vdr_taq1
vdr_taq1 <- vdr_taq1[-328]

lm_ND23_DIFF_D <- lm(ND23_DIFF ~ as.factor(vdr_taq1) + Center + Term + Gender
                     + DBP + VLDL_TG + Race, data2[-328,])
sD <- summary(lm_ND23_DIFF_D)
sD
```

```
ND23_DIFF <- ND23_DIFF_prov
vdr_taq1b <- vdr_taq1_prov

### 8.2 GANANCIA EN LAS PRUEBAS DE REPETICION MAXIMA
plot(NDRM_DIFF)

#### SELECCION DEL MODELO
NDRM_DIFF <- data2$NDRM_DIFF


lm1 <- lm(NDRM_DIFF ~  Center + Term + Gender + Age + Race + Pre.weight +
            Pre.height + pre.BMI + SBP + DBP + HDL_C +  VLDL_TG + LDL_C +  CRP +
            HOMA + Met_syn + Race, data2)

lm2<-lm(NDRM_DIFF ~ Center + Term + Gender + Age + Race + Pre.weight + Pre.height
        + pre.BMI + SBP + DBP + HDL_C +  VLDL_TG + LDL_C +  CRP +
          HOMA + Met_syn + Race, data=lm1$model)

#stepAIC(lm2, direction = 'both')

lmNDRM_DIFF <- lm(NDRM_DIFF ~ Center + Term + Gender + Age + pre.BMI + DBP + HOMA +
                    Race, data = lm1$model)
summary(lmNDRM_DIFF)

# lm_NDRM_DIFF_int <- lm(resp ~ Center + Term + Gender + Age + pre.BMI + DBP +
#FINS + HOMA + Center:Term + Center:Gender + Center:Age + Center:pre.BMI + Center:DBP
#+ Center:FINS + Center:HOMA + Term:Gender + Term:Age + Term:pre.BMI + Term:DBP +
#Term:FINS + Term:HOMA + Gender:Age + Gender:pre.BMI + Gender:DBP + Gender:FINS +
#Gender:HOMA + Age:pre.BMI + Age:DBP + Age:FINS + Age:HOMA + pre.BMI:DBP + pre.BMI:FINS
#+ pre.BMI:HOMA + DBP:FINS + DBP:HOMA + FINS:HOMA, data2)

# summary(lm_NDRM_DIFF_int)
#
# #step(lm_NDRM_DIFF_int)
#
# lm_NDRM_DIFF_int <- lm(resp ~ Center + Term + Gender + Age + pre.BMI + DBP + HOMA +
#Center:Term + Age:pre.BMI + DBP:HOMA, data2)
# summary(lm_NDRM_DIFF_int)

lm0 <- lm(NDRM_DIFF ~ Center + Term + Gender + Age + pre.BMI + DBP + HOMA + Race,
          data2)
lm1 <- lm(NDRM_DIFF ~ Term + Gender + Age + pre.BMI + DBP  + HOMA + Race, data2)
anova(lm0, lm1)

lm0 <- lm(NDRM_DIFF ~ Center + Term + Gender + Age + pre.BMI + DBP + HOMA + Race,
          data2)
lm1 <- lm(NDRM_DIFF ~ Center + Gender + Age + pre.BMI + DBP + HOMA + Race, data2)
anova(lm0, lm1)

#summary(lmNDRM_DIFF)


##### MODELO ADITIVO
valorp <- numeric()
w <- c()

for (i in 1:ncol(gendata)){
  w <- as.factor(gendata[, i])
```

```
  b <- summary(lm(NDRM_DIFF ~ as.numeric(w) + Center + Term + Gender + Age + pre.BMI
                  + DBP + Race + HOMA, data = data2))
  valorp[i] <- b$coefficients[2,4]
}

qqunif(valorp, main = 'Q-Q Plot p-valores vs. distribucion uniforme', col = 'blue4',
       las = 1, cex.lab = 0.8, cex.axis = 0.8, cex.main = 1)

par(mfrow = c(1,2))
plot(-log10(valorp), ylim=c(0, 5), main = '-log10(p-val)', pch = 20, xlab = 'SNPs',
     ylab = '-log10(pvalor)', cex.lab = 0.8, cex.main = 1, cex.axis = 0.8)
abline(h = -log10(0.05), col = 'dodgerblue3', lty = 2)
abline(h = -log10(0.05/193), col='red', lty = 2)
legend(legend = c('v. critico 0.05', 'v. critico C.Bonferroni'), col = c('dodgerblue3',
       'red'), lty = 2, 'topright', cex = 0.9)
fdr <- p.adjust(valorp, method = 'fdr')
plot(-log10(fdr),  ylim=c(0, 5), main = '-log10(p-valor) ajustados por el metodo FDR',
     pch = 20, col = 'black', cex.lab = 0.8, xlab = 'SNPs', ylab =
       '-log10(p-valor) ajustados con FDR', las = 1, cex.main = 1, cex.axis = 0.8 )
abline(h = -log10(0.05), col = 'dodgerblue3', lty = 2)
legend(legend = 'v. critico 0.05', col = 'dodgerblue3', lty = 2,
       'topright', cex = 0.85)

FDR(valorp, 0.05)

b <- colnames(gendata[,order(-log10(valorp), decreasing = TRUE)])[1:20]
b_2 <- round(sort(valorp, decreasing = FALSE), 3)[1:20]
b_b <- c("resistin a537c", "ppar gp12a", "gapd 7971637", "slc35f1 rs10484290",
         "b2b", "cav2 q130e", "akt2 2304186", "resistin c180g", "adrb2 1042713",
         "tpd52l1 3799736", "tdp52l1 9321028", "akt2 7254617", "pcr15 snp4", "pcr8 snp1",
         "c8orf68 rs6983944", "akt2 rs892118", "acdc rs1501299", "esr2 3020450",
         "tpd52l1 4896782", "akt1 c6024t c8166t")

resistin_a537c <- gendata[, b[1]]
col <- rep('dimgrey', length(NDRM_DIFF))
col[gendata[, b[1]] == 1] <- 'skyblue3'
col[gendata[, b[1]] == 2] <- 'darksalmon'
plot(NDRM_DIFF, col = col, pch = 20, lwd = 2, main = 'NDRM_DIFF segun "resistin_a537c"',
     cex.main = 1)
abline(h = mean(ND23_DIFF, na.rm = TRUE), lty = 2, col = 'red')
legend(legend = c('CC', 'CG', 'GG'), col = c('dimgrey', 'skyblue3', 'darksalmon'),
       pch = 20, 'topright', cex = 0.9)
legend(legend = 'Media', lty = 2, col = 'red', cex = 0.8, 'topleft', lwd = 2)

boxplot(NDRM_DIFF ~ gendata[, b[1]], main = 'resistin_a537c ~ NDRM_DIFF', xlab = b[1],
        col = 'lightsteelblue1', ylab = 'NDRM_DIFF', axes = FALSE, ylim = c(0,25),
        cex.main = 1, cex.lab = 0.8, cex.axis = 0.8)
axis(1, at = c(1,2,3), labels = c('AA','AC','CC'), cex.axis = 0.8)
axis(2, las = 1, cex.axis = 0.8)
points(x = 1:3,  y = (tapply(NDRM_DIFF, gendata[,b[1]], mean, na.rm = TRUE)),
       pch = 19, col = 'red', type = 'b', lty = 2)
legend(legend = 'Media', 'topright', pch = 19, col = 'red', cex = 0.9)

lm_NDRM_DIFF_A <- lm(NDRM_DIFF ~ as.numeric(resistin_a537c) + Center + Gender + Age
                     + pre.BMI + DBP + Race + HOMA,  data2)
sA <- summary(lm_NDRM_DIFF_A)
sA

par(mfrow =c (1, 2))
```

```r
plot(lm_NDRM_DIFF_A, which = 1:2, las = 1, cex.lab = 0.8, cex.axis = 0.8)

##### MODELO RECESIVO
valorp <- numeric()
wb <- c()

for (i in 1:ncol(gendata)){
  wb <- gendata[, i]
  wb[wb == 1] <- 0
  wb[wb == 2] <- 1

  p <- summary(lm(NDRM_DIFF ~ as.numeric(wb) + Center +  Gender + Age + pre.BMI
                  + DBP + Race + HOMA, data = data2))
  valorp[i] <- p$coefficients[2,4]
}
valorp <- valorp[-c(12,  27,  40,  42,  47,  93,  97, 113, 144, 145, 164)]

qqunif(valorp, main = 'Q-Q Plot p-valores vs. distribucion uniforme', col = 'blue4',
       cex.lab = 0.8, las = 1, cex.main = 1, cex.axis = 0.8)

plot(-log10(valorp), ylim=c(0, 6), main = '-log10(p-val)', pch = 20, col = 'black',
     cex.lab = 0.8, xlab = 'SNPs', ylab = '-log10(p-valor)', las = 1, cex.main = 1,
     cex.axis = 0.8)
abline(h = -log10(0.05), col = 'dodgerblue3', lty = 2)
abline(h = -log10(0.05/193), col='red', lty = 2)
legend(legend = c('v. critico 0.05', 'v. critico C.Bonferroni'), col =
         c('dodgerblue3', 'red'), lty = 2, 'topright', cex = 0.7)

b2 <- colnames(gendata[,order(-log10(valorp), decreasing = TRUE)])[12:22]
b2_2 <- round(sort(valorp, decreasing = FALSE), 3)[1:20]
b2_b <- c("akt2 2304186", "ppar gp12a", "bcl6 3774298", "tpd52l1 4896782",
          "adrb3 4994", "nos3 rs1799983", "actn3 r577x", "akt2 7254617",
          "igfbp3 6670",  "esr1 rs1042717", "adrb2 1042714")

# par(mfrow = c(1,2))
#
# col <- rep('darksalmon', length(gendata[,p2[1]]))
# col[gendata[, b2[1]] == 1] <- 'skyblue3'
# plot(NDRM_DIFF, col = col, pch = 20,  main = 'NDRM_DIFF segun "akt2_2304186"',
# ylab = 'NDRM_DIFF', las = 1, cex.lab = 0.8, xlab = 'Individuos', cex.main = 1,
# cex.axis = 0.8)
# abline(h = mean(NDRM_DIFF, na.rm = TRUE), lty = 2, col = 'red', lwd = 2)
# legend(legend = c('GG', 'GA'), col = c('skyblue3', 'darksalmon'), pch = 20,
# 'topright', cex = 0.65)
# legend(legend = 'Media', lty = 2, col = 'red', cex = 0.65, 'topleft', lwd = 2)

w <- gendata[ , b2[1]]
boxplot(NDRM_DIFF ~ gendata[, b2[1]], main = c('akt2_2304186 ~ NDRM_DIFF'),
        col = 'lightsteelblue1',
        xlab = b2[1], ylab = 'NDRM_DIFF', axes = FALSE, cex.lab = 0.8, cex.main = 1,
        ylim = c(0,25))
axis(1, at = c(1,2, 3), labels = c('GG', 'GT', 'TT'), cex.axis = 0.8)
axis(2, las = 1, cex.axis = 0.8)
points(x = 1:3,  y = (tapply(NDRM_DIFF, gendata[,b2[1]], mean,
                             na.rm = TRUE)),  pch = 19, col = 'red', type = 'b',
       lty = 2)
legend(legend = 'Media', 'topright', pch = 19, col = 'red', cex = 0.8)

akt2_2304186b <- gendata[, 'akt2_2304186']
```

```
akt2_2304186b[gendata[, 'akt2_2304186'] == 1] <- 0
akt2_2304186b[gendata[, 'akt2_2304186'] == 2] <- 1

lm_NDRM_DIFF_B <- lm(NDRM_DIFF ~ as.numeric(akt2_2304186b) + Center + Term + Gender
                     + DBP + VLDL_TG + Race,  data2)
sB <- summary(lm_NDRM_DIFF_B)
sB

par(mfrow =c (1, 2))
plot(lm_NDRM_DIFF_B, which = 1:2, las = 1, cex.lab = 0.8, cex.axis = 0.8)


##### MODELO DOMINANTE
valorp <- numeric()

for (i in 1:ncol(gendata)){
  wc <- gendata[, i]
  wc[wc == 2] <- 1

  p <- summary(lm(NDRM_DIFF ~ as.numeric(wc) + Age + Center + Term + Gender + DBP +
                    Race + HOMA + pre.BMI,  data = data2))
  valorp[i] <- p$coefficients[2,4]
}

qqunif(valorp, main = 'Q-Q␣Plot␣p-valores␣vs.␣distribucion␣uniforme', col = 'blue4',
       cex.lab = 0.8, las = 1, cex.main = 1, cex.axis = 0.8)

par(mfrow = c(1,2))
plot(-log10(valorp), ylim=c(0, 5), main = '-log10(p-val)', pch = 20, col = 'black',
     cex.lab = 0.8, xlab = 'Individuos', ylab = '-log10(p-valor)', las = 1, cex.main = 1,
     cex.axis = 0.8)
abline(h = -log10(0.05), col = 'dodgerblue3', lty = 2)
abline(h = -log10(0.05/193), col='red', lty = 2)
legend(legend = c('v.␣critico␣0.05', 'v.␣critico␣C.Bonferroni'), col = c('dodgerblue3',
      'red'), lty = 2, 'topright', cex = 0.9)
fdr <- p.adjust(valorp, method = 'fdr')
plot(-log10(fdr),  ylim=c(0, 5), main = '-log10(p-valor)␣ajustados␣por␣el␣metodo␣FDR',
     pch = 20, col = 'black', cex.lab = 0.8, xlab = 'SNPs', ylab =
       '-log10(p-valor)␣ajustados␣con␣FDR', las = 1, cex.main = 1, cex.axis = 0.8 )
abline(h = -log10(0.05), col = 'dodgerblue3', lty = 2)
legend(legend = 'v.␣critico␣0.05', col = 'dodgerblue3', lty = 2,
       'topright', cex = 0.85)

FDR(valorp, 0.05)

b3 <- colnames(gendata[,order(-log10(valorp), decreasing = TRUE)])[1:10]
b3_2 <- round(sort(valorp, decreasing = FALSE), 3)[1:10]
b3_b <- c("b2b", "adrb2␣1042713", "resistin␣a537c", "gapd␣7971637", "cav2␣q130e",
          "slc35f1␣rs10484290", "pcr15␣snp4", "resistin␣c180g", "ppar␣gp12a",
          "mgst3␣4147542")

# par(mfrow = c(1,2))
#
# col <- rep('dimgrey', length(NDRM_DIFF))
# col[gendata[, p3[1]] == 1] <- 'skyblue3'
# col[gendata[, p3[1]] == 2] <- 'darksalmon'
# plot(NDRM_DIFF, col = col, pch = 20, lwd = 2, main = 'NDRM_DIFF segun "b2b"',
#cex.main = 1,
# las = 1, cex.main = 1, cex.lab = 0.9, cex.axis = 0.9)
```

```
# abline(h = mean(NDRM_DIFF, na.rm = TRUE), lty = 2, col = 'red')
# legend(legend = c('CC', 'TC', 'TT'), col = c('dimgrey', 'skyblue3', 'darksalmon'),
# pch = 20, 'topright', cex = 0.9)
# legend(legend = 'Media', lty = 2, col = 'red', cex = 0.8, 'topleft', lwd = 2)

boxplot(NDRM_DIFF ~ gendata[, b3[1]], main = c('b2b ~ NDRM_DIFF'), xlab = b3[1],
        ylab = 'NDRM_DIFF', axes = FALSE, cex.main = 1, cex.lab = 0.8, cex.axis = 0.8,
        col = 'lightsteelblue1')
axis(1, at = c(1,2,3), labels = c('CC','TC','TT'), cex.axis = 0.8)
axis(2, las = 1, cex.axis = 0.8)
points(x = 1:3,  y = (tapply(NDRM_DIFF, gendata[,b3[1]], mean, na.rm = TRUE)),
       pch = 19, col = 'red', type = 'b', lty = 2)
legend(legend = 'Media', 'topright', pch = 19, col = 'red', cex = 0.9)

b2bc <- gendata[, 'b2b']
b2bc[gendata[, 'b2b'] == 2] <- 1

lm_NDRM_DIFF_C <- lm(NDRM_DIFF ~ as.numeric(b2bc) + Age + Center + Term + Gender + DBP
                     + Race + HOMA + pre.BMI, data2)
sC <- summary(lm_NDRM_DIFF_C)
sC

par(mfrow =c (1, 2))
plot(lm_NDRM_DIFF_C, which = 1:2, las = 1, cex.lab = 0.8, cex.axis = 0.8)


##### MODELO CODOMINANTE
valorp1 <- numeric()
valorp2 <- numeric()
w <- c()

for (i in 1:ncol(gendata)){
  w <- as.factor(gendata[, i])
  lm0 <- lm(NDRM_DIFF ~ w + Center + Gender + Age + pre.BMI + DBP + Race + HOMA,
            data = data2)
  lm1 <- lm(NDRM_DIFF ~ Center +  Gender + Age + pre.BMI + DBP + Race + HOMA,
            data = data2)
  a <- anova(lm0,lm1)
  valorp1 <- c(valorp1, a$`Pr(>F)`[2])
}

#par(mfrow = c(1, 2))
qqunif(valorp1, main = 'Q-Q Plot p-valores vs. dist uniforme', col = 'blue4',
       cex.lab = 0.8, las = 1, cex.axis = 0.8)

par(mfrow = c(1,2))
plot(-log10(valorp1), ylim=c(0, 5), main = '-log10(p-valor)', pch = 20, xlab = 'SNPs',
     ylab = '-log10(p-valor)', las = 1, cex.lab = 0.8, cex.axis = 0.8, cex.main = 1)
abline(h = -log10(0.05), col = 'dodgerblue3', lty = 2)
abline(h = -log10(0.05/193), col='red', lty = 2)
legend(legend = c('v. critico 0.05', 'v. critico C.Bonferroni'), col =
         c('dodgerblue3', 'red'), lty = 2,  'topright', cex = 0.9)
fdr <- p.adjust(valorp1, method = 'fdr')
plot(-log10(fdr),  ylim=c(0, 5), main = '-log10(p-valor) ajustados por el metodo FDR',
     pch = 20, col = 'black', cex.lab = 0.8, xlab = 'SNPs', ylab =
       '-log10(p-valor) ajustados con FDR', las = 1, cex.main = 1, cex.axis = 0.8 )
abline(h = -log10(0.05), col = 'dodgerblue3', lty = 2)
legend(legend = 'v. critico 0.05', col = 'dodgerblue3', lty = 2,
       'topright', cex = 0.85)
```

```
FDR(valorp, 0.05)

b4 <- colnames(gendata[,order(-log10(valorp1), decreasing = TRUE)])[1:10]
b4_2 <- round(sort(valorp1, decreasing = FALSE)[1:12], 3)
b4_b <- c("b2b", "il15ra␣3136618", "adrb2␣1042713", "ppar␣gp12a", "resistin␣a537c",
          "ankrd6␣a550t", "pcr15␣snp1", "akt2␣2304186", "gapd␣7971637",
          "slc35f1␣rs10484290")

# par(mfrow = c(1,2))
#
# col <- rep('dimgrey', length(NDRM_DIFF))
# col[gendata[, p3[1]] == 1] <- 'skyblue3'
# col[gendata[, p3[1]] == 2] <- 'darksalmon'
# plot(NDRM_DIFF, col = col, pch = 20, lwd = 2, main = 'NDRM_DIFF segun "b2b"',
#cex.main = 1, las = 1, cex.main = 1,
#      cex.lab = 0.9, cex.axis = 0.9)
# abline(h = mean(NDRM_DIFF, na.rm = TRUE), lty = 2, col = 'red')
# legend(legend = c('CC', 'TC', 'TT'), col = c('dimgrey', 'skyblue3', 'darksalmon'),
#pch = 20, 'topright', cex = 0.9)
# legend(legend = 'Media', lty = 2, col = 'red', cex = 0.8, 'topleft', lwd = 2)

boxplot(NDRM_DIFF ~ gendata[, b3[1]], main = c('b2b␣~␣NDRM_DIFF'), xlab = b3[1],
        ylab = 'NDRM_DIFF', axes = FALSE, cex.main = 1, cex.lab = 0.8, cex.axis = 0.8,
        col = 'lightsteelblue1')
axis(1, at = c(1,2,3), labels = c('CC','TC','TT'), cex.axis = 0.8)
axis(2, las = 1, cex.axis = 0.8)
points(x = 1:3,  y = (tapply(NDRM_DIFF, gendata[,b3[1]], mean, na.rm = TRUE)),
       pch = 19, col = 'red', type = 'b', lty = 2)
legend(legend = 'Media', 'topright', pch = 19, col = 'red', cex = 0.9)

b2bd <- gendata[ , b4[1]]
b2bd[b2bd == 0] <- 'CC'
b2bd[b2bd == 1] <- 'TC'
b2bd[b2bd == 2] <- 'TT'
b2bd <- as.factor(b2bd)

lm_NDRM_DIFF_D <- lm(NDRM_DIFF ~ as.factor(b2bd) + Center +  Gender + Age + pre.BMI +
                       DBP + Race + HOMA,  data2)
sD <- summary(lm_NDRM_DIFF_D)
sD

par(mfrow =c (1, 2))
plot(lm_NDRM_DIFF_D, which = 1:2, las = 1, cex.lab = 0.8, cex.axis = 0.8)

### ESTIMACION DE HAPLOTIPOS
dhapl <- rawdata[rownames(gendata), ]
dhapl <- dhapl[which(data2[, 'Race'] == 'Caucasian'), ]
dhapl <- dhapl[, which(names(dhapl)==c("resistin_c30t", "resistin_c398t", "resistin_g540a",
                                       "resistin_c980g", "resistin_c180g", "resistin_a537c"))]


set.seed(2024)
for(i in 1:ncol(dhapl[,1:2])){
  n2 <- sum(dhapl[, i] == 'CC', na.rm = TRUE)
  n1 <- sum(dhapl[, i] == 'CT', na.rm = TRUE)
  n0 <- sum(dhapl[, i] == 'TT', na.rm = TRUE)
  ntot <- n2 + n1 + n0
```

```
  p2 <- n2 / ntot
  p1 <- n1 / ntot
  p0 <- n0 / ntot

  nmis <- sum(is.na(dhapl[, i]))

  pseudo <-sample(c('CC', 'CT', 'TT'), nmis, replace = TRUE, prob = c(p2, p1, p0))

  dhapl[, i][is.na(dhapl[, i])] <- pseudo
}


set.seed(2024)
i <- 3
n2 <- sum(dhapl[, i] == 'GG', na.rm = TRUE)
n1 <- sum(dhapl[, i] == 'GA', na.rm = TRUE)
n0 <- sum(dhapl[, i] == 'AA', na.rm = TRUE)
ntot <- n2 + n1 + n0

p2 <- n2 / ntot
p1 <- n1 / ntot
p0 <- n0 / ntot

nmis <- sum(is.na(dhapl[, i]))

pseudo <-sample(c('GG', 'GA', 'AA'), nmis, replace = TRUE, prob = c(p2, p1, p0))

dhapl[, i][is.na(dhapl[, i])] <- pseudo



set.seed(2024)
for(i in 4:5){
  n2 <- sum(dhapl[, i] == 'GG', na.rm = TRUE)
  n1 <- sum(dhapl[, i] == 'CG', na.rm = TRUE)
  n0 <- sum(dhapl[, i] == 'CC', na.rm = TRUE)
  ntot <- n2 + n1 + n0

  p2 <- n2 / ntot
  p1 <- n1 / ntot
  p0 <- n0 / ntot

  nmis <- sum(is.na(dhapl[, i]))

  pseudo <-sample(c('GG', 'CG', 'CC'), nmis, replace = TRUE, prob = c(p2, p1, p0))

  dhapl[, i][is.na(dhapl[, i])] <- pseudo
}

set.seed(2024)
i <- 6
n2 <- sum(dhapl[, i] == 'AA', na.rm = TRUE)
n1 <- sum(dhapl[, i] == 'AC', na.rm = TRUE)
n0 <- sum(dhapl[, i] == 'CC', na.rm = TRUE)
ntot <- n2 + n1 + n0

p2 <- n2 / ntot
p1 <- n1 / ntot
p0 <- n0 / ntot
```

```
nmis <- sum(is.na(dhapl[, i]))

pseudo <-sample(c('AA', 'AC', 'CC'), nmis, replace = TRUE, prob = c(p2, p1, p0))

dhapl[, i][is.na(dhapl[, i])] <- pseudo


snp1 <- dhapl[, 1]
snp2 <- dhapl[, 2]
snp3 <- dhapl[, 3]
snp4 <- dhapl[, 4]
snp5 <- dhapl[, 5]
snp6 <- dhapl[, 6]

Geno <- cbind(substr(snp1, 1, 1), substr(snp1, 2, 2),
              substr(snp2, 1, 1), substr(snp2, 2, 2),
              substr(snp3, 1, 1), substr(snp3, 2, 2),
              substr(snp4, 1, 1), substr(snp4, 2, 2),
              substr(snp5, 1, 1), substr(snp5, 2, 2),
              substr(snp6, 1, 1), substr(snp6, 2, 2))

snpnames <- c("snp1", "snp2","snp3","snp4","snp5","snp6")

set.seed(2024)
HaploEM <- haplo.em(Geno, locus.label = snpnames, control = haplo.em.control
                    (min.posterior = 0.005))
HaploEM


s <- summary(HaploEM, show.haplo = TRUE)

s[1:6, ]

vhapl <- c()
i <- 1
while(i <= (length(s$subj.id)-1)){
  if(s$subj.id[i]==s$subj.id[i+1]){
    j <- i
    n <- 1
    p <- s$posterior[j]
    while(s$subj.id[j]==s$subj.id[j+1]){
      n <- n + 1
      j <- j + 1
      p <- c(p, s$posterior[j])
    }
    vhapl <- c(vhapl, paste(s$hap1.snp1[i + which.max(p) - 1],
                            s$hap1.snp2[i +  which.max(p) - 1],
                            s$hap1.snp3[i + which.max(p) - 1],
                            s$hap1.snp4[i + which.max(p) - 1],
                            s$hap1.snp5[i + which.max(p) - 1],
                            s$hap1.snp6[i + which.max(p) - 1],
                            s$hap2.snp1[i + which.max(p) - 1],
                            s$hap2.snp2[i + which.max(p) - 1],
                            s$hap2.snp3[i + which.max(p) - 1],
                            s$hap2.snp4[i + which.max(p) - 1],
                            s$hap2.snp5[i + which.max(p) - 1],
                            s$hap2.snp6[i + which.max(p) - 1]))
    i <- i + n
```

```
  }else{
    vhapl <- c(vhapl, paste(s$hap1.snp1[i], s$hap1.snp2[i], s$hap1.snp3[i],
                            s$hap1.snp4[i], s$hap1.snp5[i], s$hap1.snp6[i],
                            s$hap2.snp1[i], s$hap2.snp2[i], s$hap2.snp3[i],
                            s$hap2.snp4[i], s$hap2.snp5[i], s$hap2.snp6[i]))
    i <- i + 1
  }
}


if(s$subj.id[i]==s$subj.id[i-1]){
  vhapl <- vhapl
}else{
  vhapl <- c(vhapl, paste(s$hap1.snp1[i], s$hap1.snp2[i], s$hap1.snp3[i],
                          s$hap1.snp4[i], s$hap1.snp5[i], s$hap1.snp6[i],
                          s$hap2.snp1[i], s$hap2.snp2[i], s$hap2.snp3[i],
                          s$hap2.snp4[i], s$hap2.snp5[i], s$hap2.snp6[i]))
}

orvhapl <- c()

for(i in 1:length(vhapl)){
  orvhapl <- c(orvhapl, paste(strsplit(vhapl[i], ' ')[[1]], collapse = ''))
}

### ELABORACION DEL MODELO
#### GANANCIA DE FUERZA ISOMETRICA
dcau <- data2[which(data2[, 'Race'] == 'Caucasian'), ]

orvhapl2 <- rep("Otros", length(orvhapl))
orvhapl2[orvhapl=='CCGCCACCGGCA'] <- 'CCGCCA/CCGGCA'
orvhapl2[orvhapl=='CCGGCACTACGA'] <- 'CCGGCA/CTACGA'
orvhapl2[orvhapl=='CCGGCACCGGCA'] <- 'CCGGCA/CCGGCA'
orvhapl2[orvhapl=='CCGCCACTACGA'] <- 'CCGCCA/CTACGA'
orvhapl2 <- as.factor(orvhapl2)

ND23_DIFF_cau <- dcau$ND23_DIFF

boxplot(ND23_DIFF_cau ~ orvhapl2, main = 'Diplotipos mas comunes vs. ganancia de
        fuerza isometrica', xlab = 'Diplotipos',
        ylab = 'ND23_DIFF',  cex.lab = 0.8, cex.main = 1, las = 1, cex.axis = 0.75,
        col = 'lightsteelblue1')
points(x = 1:5,  y = (tapply(ND23_DIFF_cau, orvhapl2, mean, na.rm = TRUE)),
       pch = 19, col = 'red', type = 'b', lty = 2)
legend(legend = 'Media', 'topright', cex = 0.8, pch = 19, col = 'red')

lmhap1 <- lm(ND23_DIFF_cau ~  Center + Term + Gender + DBP + VLDL_TG + Age +
               pre.BMI + HOMA + as.factor(orvhapl2), dcau)
#car::Anova(lmhap1, type = 'III')

lmhap1 <- lm(ND23_DIFF_cau ~  Center + Term + Gender + DBP + VLDL_TG +
               as.factor(orvhapl2), dcau)
car::Anova(lmhap1, type = 'III')

summary(lmhap1)


lmhap2 <- lm(ND23_DIFF_cau ~  Center + Term + Gender + DBP + VLDL_TG + Age + pre.BMI +
               HOMA + as.factor(orvhapl), dcau)
```

```
#car::Anova(lmhap2, type = 'III')

lmhap2 <- lm(ND23_DIFF_cau ~  Center + Term + Gender + DBP +  as.factor(orvhapl), dcau)
car::Anova(lmhap2, type = 'III')

summary(lmhap2)

df <- data.frame()
for(i in 1:length(orvhapl)){
  df[i,1] <- substr(orvhapl[i], 1, 6)
  df[i,2] <- substr(orvhapl[i], 7, 12)
}

CTACGA <- rep(1, length(orvhapl))
CTACGA[df[, 1] == 'CTACGA'] <- 0
CTACGA[df[, 2] == 'CTACGA'] <- 0
CTACGA <- as.factor(CTACGA)

table(CTACGA)

lmhap3 <- lm(ND23_DIFF_cau ~ Center + Term + Gender + DBP + VLDL_TG + Age + pre.BMI +
               HOMA + as.factor(CTACGA), dcau)
#car::Anova(lmhap3, type = 'III')

lmhap3 <- lm(ND23_DIFF_cau ~ Center + Term + Gender + DBP + VLDL_TG + as.factor(CTACGA),
             dcau)
car::Anova(lmhap3, type = 'III')

#### GANANCIA DEL TEST DE REPETICION MAXIMA
NDRM_DIFF_cau <- dcau$NDRM_DIFF
boxplot(NDRM_DIFF_cau~orvhapl2, main = 'Diplotipos␣mas␣comunes␣vs.␣ganancia
␣␣␣␣␣␣␣␣test␣repeticion␣maxima', xlab = 'Diplotipos',
        ylab = 'NDRM_DIFF',  cex.lab = 0.8, cex.main = 1, las = 1, cex.axis = 0.75,
        col = 'lightsteelblue1')
points(x = 1:5,  y = (tapply(NDRM_DIFF_cau, orvhapl2, mean, na.rm = TRUE)),
       pch = 19, col = 'red', type = 'b', lty = 2)
legend(legend = 'Media', 'topright', cex = 0.7, pch = 19, col = 'red')

lmhap4 <- lm(NDRM_DIFF_cau ~  Center + Term + Gender + DBP + VLDL_TG + Age + pre.BMI
             + HOMA + as.factor(orvhapl2), dcau)
#car::Anova(lmhap4, type = 'III')

lmhap4 <- lm(NDRM_DIFF_cau ~Center + Term + Gender + DBP +  pre.BMI+ as.factor(orvhapl2),
             dcau)
car::Anova(lmhap4, type = 'III')


lmhap5 <- lm(NDRM_DIFF_cau ~  Center + Term + Gender + DBP + VLDL_TG + Age + pre.BMI
             + HOMA + as.factor(orvhapl), dcau)
#car::Anova(lmhap5, type = 'III')

lmhap5 <- lm(NDRM_DIFF_cau ~ Center +  Gender + DBP +  Age + pre.BMI + as.factor(orvhapl),
             dcau)
car::Anova(lmhap5, type = 'III')


summary(lmhap5)

#### MODELO MULTIVARIANTE
```

```
resp <- cbind(dcau$ND23_DIFF, dcau$NDRM_DIFF)

lmhap6 <- lm(resp ~ Center + Term + Gender + DBP + VLDL_TG + Age + pre.BMI + HOMA +
               as.factor(orvhapl2), dcau)
#car::Anova(lmhap6, type = 'III')

#lmhap6 <- lm(resp ~ Center + Term + Gender + DBP + VLDL_TG + Age + pre.BMI + HOMA +
#as.factor(orvhapl)+ as.factor(orvhapl):Center + as.factor(orvhapl):Term+
#as.factor(orvhapl):Gender+ as.factor(orvhapl):DBP+ as.factor(orvhapl):VLDL_TG+
#as.factor(orvhapl):Age + as.factor(orvhapl):pre.BMI+ as.factor(orvhapl):HOMA, dcau)
#car::Anova(lmhap, type = 'III')

lmhap6 <- lm(resp ~ Center + Term + Gender + DBP +  as.factor(orvhapl2), dcau)
#summary(lmhap6)
car::Anova(lmhap6, type = 'III')


resp <- cbind(dcau$ND23_DIFF, dcau$NDRM_DIFF)

lmhap <- lm(resp ~ Center + Term + Gender + DBP + VLDL_TG + Age + pre.BMI + HOMA +
              as.factor(orvhapl), dcau)
#car::Anova(lmhap, type = 'III')

#lmhap <- lm(resp ~ Center + Term + Gender + DBP + VLDL_TG + Age + pre.BMI + HOMA +
#as.factor(orvhapl)+ as.factor(orvhapl):Center + as.factor(orvhapl):Term+
#as.factor(orvhapl):Gender+ as.factor(orvhapl):DBP+ as.factor(orvhapl):VLDL_TG +
#as.factor(orvhapl):Age + as.factor(orvhapl):pre.BMI+ as.factor(orvhapl):HOMA, dcau)
#car::Anova(lmhap, type = 'III')

lmhap <- lm(resp ~ Center + Term + Gender + DBP + as.factor(orvhapl), dcau)
#summary(lmhap)
car::Anova(lmhap, type = 'III')

summary(lmhap)
```